Mining the genomes of gliding bacteria with emphasis on *Herpetosiphon aurantiacus*

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Lavanya Natesan

aus

Mettupalayam, India

Bonn 2012

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachterin : Prof. Dr. G. M. König
2. Gutachterin : Prof. Dr. E. Kostenis

Tag der Promotion : 20.11.2012
Erscheinungsjahr: 2012

Publications that was done during this PhD but not included in this thesis

C. Almeida, Y. Hemberger, S.M. Schmitt, S. Bouhired, L. Natesan, S. Kehraus, K. Dimas, M. Gütschow, G. Bringmann, G.M. König; Marilines A–C: Novel Phthalimidines from the Sponge-Derived Fungus *Stachylidium* sp. *Chem. Eur. J.* **18**, 8827–8834 (2012).

S. Kastner, S. Müller, L. Natesan, G.M. König, R. Guthke, M. Nett; 4-Hydroxyphenylglycine biosynthesis in *Herpetosiphon aurantiacus*: a case of gene duplication and catalytic divergence. *Archives of Microbiology* **194**, 557–566 (2012).

M.F. Elsebai, L.Natesan, S. Kehraus, I.E. Mohamed, G. Schnakenburg, F. Sasse, S. Shaaban, M. Gütschow, G.M. König; HLE-inhibitory alkaloids with a polyketide skeleton from the marine-derived fungus *Coniothyrium cereale. J. Nat. Prod.* **74**, 2282–2285 (2011).

Research Presentations

L. Natesan, M. Nett, G.M. König; MbtH-like proteins, a member of the nonribosomal peptide biosynthetic machinery?; Abstract accepted at the International VAAM workshop on Biology of Bacteria producing Natural Products, Bonn, Germany, Sep 28-30, 2011.

L. Natesan, G.M König; The world of nonribosomal peptides, Oral presentations at the workshops conducted by Biotech-Pharma on Biotechnology aspects of natural products, Bonn, Germany, Oct 12-14, 2009 and Nov 9-11, 2011.

L. Natesan, M. Nett, G.M. König; Genetically encoded biosynthetic potential of *Herpetosiphon aurantiacus*; Poster and Oral presentation at the International VAAM workshop on Biology of Bacteria producing Natural Products, Tübingen, Germany, Sep 26-28, 2010.

L. Natesan, M. Nett, G.M. König; Towards the isolation of bio-active natural products, deciphered from the *Herpetosiphon aurantiacus* genome sequence, Poster presentation at the Royal Society of Chemistry (RSC) on Directed biosynthesis II: Discovery, Evolution Function, Durham, United Kingdom, Sep 15-17, 2010.

L. Natesan, M. Nett, G.M. König; Genomics-based approach towards the discovery of lipopeptide from *Herpetosiphon aurantiacus*; Poster presentation at the European Science Foundation (ESF) Research Congress on Natural Products Chemistry, Biology and Medicine III, Acquafredda di Maratea, Italy, Sep 5-10, 2010.

L. Natesan, M. Nett, G.M. König; Genomics based discovery of natural products from *Herpetosiphon aurantiacus*; Poster presentation at the 58th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Product Research, Berlin, Germany, Aug 29-Sep 2, 2010.

L. Natesan, M. Nett, G.M König; Genomics driven discovery of bioactive compounds from *Herpetosiphon aurantiacus*, Oral presentations at Central Colloquium of the Graduate College (GRK677), Bonn, Germany, Dec 8, 2008 and Jan 18, 2010.

L. Natesan, M. Nett, G.M König; Characterisation of an unusual substrate, recognized by a nonribosomal peptide synthetase gene in *Herpetosiphon aurantiacus*, Poster and Oral presentation at the International Symposium conducted by Central Colloquium of the Graduate College (GRK677), Bonn, Germany, Sep 21-23, 2009.

L. Natesan, M. Nett, G.M König; Genomics driven drug discovery from *Herpetosiphon aurantiacus*, Poster presentation at the Minisymposium conducted by Central Colloquium of the Graduate College (GRK677), Bonn, Germany, Apr-1, 2009.

L. Natesan, G.M König; The world of nonribosomal peptides, Oral presentation at Central Colloquium of the Graduate College (GRK677), Bonn, Germany, Jun 2, 2008.

Acknowledgement

Words fail me to express how much I am indebted to my supervisor, Prof. G.M. Konig for giving me the opportunity to conduct this research. She provided invaluable support and unflinching encouragement throughout this period. My mere expression of thanks likewise does not suffice for her patience during all the meetings, discussions and continuous interest in supervising me.

Big thanks to Prof. Dr. E. Kostenis for the co-examination of my thesis and Prof. Dr. D. Imhof and Prof. Dr. H. Häberlein for serving on my examination committee.

This research was undertaken whilst in receipt of a scholarship from Deustche Forschungsgemeinschaft under GRK 677, and the thesis was written whilst employed as a researcher at the University of Bonn.

I am most grateful to Dr. M. Nett, who laid foundation for this project and provided the opportunity for me to work as a Visiting Scientist in his lab in Leibniz Institute for Natural Product Research and Infection Biology. Also, for being very supportive right from the beginning of this work and for sharing knowledge about heterologous protein expression and for constant collaboration throughout this research period.

Big appreciation to Dr. S. Kehraus, who helped me with the analytical methods and to Dr. Max Crüsemaan for running assays.

It is great pleasure to acknowledge Dr.T. Schäberle for helping me on carrying out protein assay and for related discussions.

I express my sincere thanks to Ms. E. Eugereva and Ms. E. Neu for helping me on LC-MS measurements and initial cultivation process and for being very friendly always and for socialising with me right from the beginning.

I convey special acknowledgement to E. Goralski in the S1 lab for helping me with procuring the materials required for the research.

I am overwhelmed with gratitude towards Dr. C. wagner in S1 lab for introducing me to the molecular lab.

I thank Dr.T. Höver for sharing the information about his transformation experiments.

It is a pleasure to work with S. Bouhired who has been affable and kind in providing valuable suggestions and discussions on my research work and also for my life in Germany.

Abbreviations

ν	wave number [cm^{-1}]
λ	wavelength [nm]
ACP	Acyl carrier protein
A domain	Adenylation domain
AT	Acyl transferase
bp	Base pair
C domain	Condensation domain
HC domain	Cyclisation domain
CoA	Coenzyme A
Da	Dalton
DAHP	3-deoxy-D-arabino-heptulosonate 7-phosphate
DH	Dehydratase
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	N,N,N´,N´-Ethylendiamintetraacetat
ER	Enoylreductase
FAS	Fatty acid synthesis
IPTG	Isopropyl-β-D-thiogalactosid
kb	Kilobases
kDa	Kilodalton
KR	Ketoreductase
KS	Ketosynthase
MALDI	Matrix-assisted laser desorption/ionisation
mRNA	Messenger RNA
NADH	Nicotinamide adenine dinucleotide
NRPS	Non-ribosomal peptide synthetase
OD	Optical density
MO	Monooxygenase
O-MT	*O*-methyltransferase
abs	absorbance
ATP	adenosine triphosphate

CC	column chromatography
dNTP	desoxynucleoside triphosphate
DSM	Deutsche Stammsammlung für Mikroorganismen und Zellkulturen – strain accession number
E. coli	*Escherichia coli*
ESI	electro spray ionisation
EtOAc	ethyl acetate
EtOH	ethanol
HLE	human leukocyte elastase
HPLC	high performance liquid chromatography
HR	high resolution
hr	hour
Hz	Hertz
LC	liquid chromatography
MeOH	methanol
MHz	Megahertz
min	minute
MRSA	methicillin resistant *Staphylococcus aureus*
MS	mass spectrometry
NMR	nuclear magnetic resonance
NRPS	nonribosomal peptide synthetase
^{o}C	degrees Celsius
OR	Oxidoreductase
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCP	Peptidyl carrier protein
PEP	phosphoenolpyruvate
PPi	Inorganic pyrophosphate
RNA	Ribonucleic acid
RT-PCT	Reversed transcription-PCR
SAM	*S*-adenosylmethionine
SDS	Sodium dodecylsulfate
TE	Thioesterase

Me	methyl
mg	Milligram
g	gram
ml	millilitre
L	litre
min	Minute
M	molar
mM	millimolar
µg	microgram
µl	microlitre
TEMED	N,N,N',N'-Tetramethylethylendiamin
PCR	polymerase chain reaction
PKS	polyketide synthase
ppm	parts per million
RI	refractive index
RNA	ribonucleic acid
RP	reversed phase
rpm	rounds per minute
RT	room temperature
sp.	species
UV	ultraviolet
VLC	vacuum-liquid chromatography

Amino acid	3 letter code	1 letter code	Amino acid	3 letter code	1 letter code
Glycine	gly	G	Threonine	thr	T
Alanine	ala	A	Cysteine	cys	C
Valine	val	V	Tyrosine	tyr	Y
Leucine	leu	L	Asparagine	asn	N
Isoleucine	ile	I	Glutamine	gln	Q
Methionine	met	M	Aspartate	asp	D
Phenylalanine	phe	F	Glutamate	glu	E
Tryptophan	trp	W	Lysine	lys	K
Proline	pro	P	Arginine	arg	R
Serine	ser	S	Histidine	his	H

Table of contents

Tables

Figures

1. Introduction

1.1 Gliding bacteria

Motility is an important characteristic of bacteria that enables them to move towards or away from attractants or repellants, respectively. This facilitates bacteria to choose suitable conditions for their living. Bacterial motility is of three types: flagella, twitching and gliding. Among them, the gliding motility is fascinating because of its distinct mechanisms encountered in different bacteria (McBride, 2001, Jarrell and McBride, 2008). Bacteria which move by gliding on the surface have been isolated from almost all environment including soil, marine samples, sewage plants and hydrothermal vents. Till now, no similarity has been observed among these bacteria except their pattern of locomotion. In 2007, Nett and König clustered gliding bacteria based on their phylogenetic origin (Table 1-1). It was apparent from the grouping that only the phylum Chloroflexi exclusively contains gliding bacteria, whereas other phyla include both gliding and non-gliding species (Nett and König, 2007).

Phylum	Genera
Chloroflexi	***Chloroflexus**, Chloronema, Heliothrix, Oscillothrix, **Herpetosiphon***
Cyanobacteria	*Chroococcus, Gloeocapsa, **Synechocystis**, Dermocarpella, Staniera, Myxosarcina, Pleurocapsa, Arthrospira, **Geitlerinema**, Leptolyngbya, **Lyngbya**, , **Oscillatoria**, **Plankothrix**, Pseudoanabaena, Spirulina, **Symploca**, **Microcoleus**, Trichodesmium, **Tychonema**, **Anabaena**, **Cylindrospermum**, **Nodularia**, **Nostoc**, **Calothrix**, Calothrix, **Rivularia**, **Tolypothrix**, **Chlorogloeopsis**, **Fischerella***
Chlorobi	*Chloroherpeton*
Proteobacteria	*Thithrix, Achromatium, Beggiatoa, Leucothrix, Thiploca, **Lysobacter**, **Myxococcus**, **Corallococcus**, Pyxicoccus, **Cystobacter**, **Archangium**, Hyalangium, **Melittangium**, **Stigmatella**, Polyangium, **Byssophaga**, **Chondromyces**, Hapolangium, Jahnia, **Sorangium**, **Nannocystis**, Kofleria*
Firmicutes	*Heliobacterium, Mycoplasma, Filibacter*
Planctomycetes	*Isosphaera*
Bacteroidetes	*Marinilabila, **Flavobacterium**, Capnocytophaga, Cellulophaga, Gelidibacter, Psychroflexus, Tenacibaculum, Zobellia, Sphingobacterium, Pedobacter, **Saprospira**, Haliscomenobacte, Lewinella, **Flexibacter**, **Cytophaga**, Microscilla, Runella, Spirosoma, Sporocytophaga, **Flexithrix**, Persicobacter, Thermonema, Chtinophaga, Toxothrix*

Table 1-1: Grouping of gliding bacteria based on their phylogenetic origin. Genera known as the producers of secondary metabolites are highlighted in bold. This table is reproduced from Nett and König, 2007.

1.1.1 Secondary metabolites of gliding bacteria

Gliding bacteria are a prolific source of natural products, the latter being extremely valuable in drug discovery research, where novel lead compounds are urgently needed (Newman and Cragg, 2012). Myxobacteria and cyanobacteria with their secondary metabolic profile comparable to that of the actinomycetes are the most promising groups of gliding bacteria in this respect. From myxobacteria, compounds with around 100 different core structures and 500 derivatives thereof have been isolated (Weissman and Müller, 2009; Weissman and Müller, 2010). In this context it is important to note, that a semi-synthetic derivative of the myxobacterial metabolite epothilone (*Sorangium cellulosum*), ixabepilone, has been approved for the treatment of metastatic breast cancer (Vahdat, 2008). Indeed, the drug was proven to be effective on cancer cells resistant to chemotherapy with other agents. Similar to myxobacteria, more than 800 compounds with a wide range of activities from cytotoxic to antimicrobial were isolated from cyanobacteria (Tan, 2007; Wagoner et al., 2007). Secondary metabolites with intriguing structures were isolated from other groups of gliding bacteria as well including some genera of the phylum Bacteroidetes (Nett and König, 2007). In the latter cases however, the proportion is low when compared to myxobacteria and cyanobacteria.

Overall, the investigation of gliding bacteria resulted in the isolation of new metabolites with striking structural and biological properties (Figure 1-1). Many of the compounds that have been isolated belong to a class of compounds called polyketides and nonribosomal peptides.
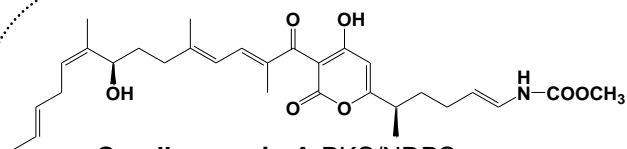
Corallopyronin A-PKS/NRPS
(*RNA Polymerase inhibitor*)
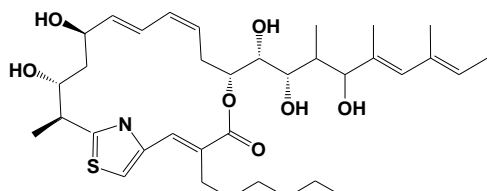
Thuggacin A-PKS/NRPS
(*Cytochrome oxidase inhibitor*)
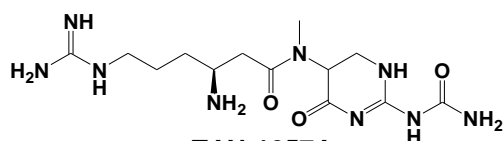
TAN-1057A
(*Protein synthesis inhibitor*)
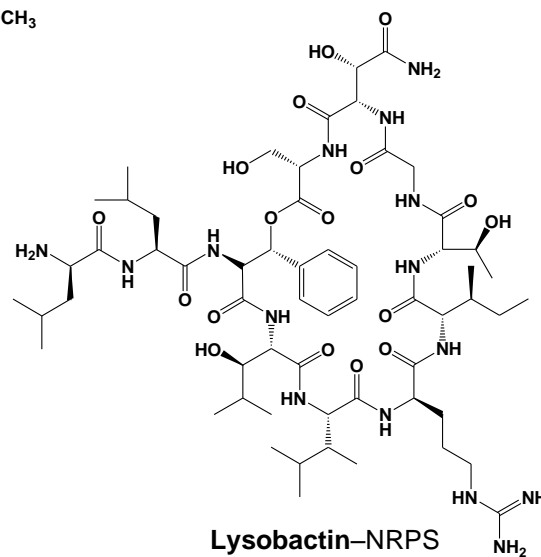
Lysobactin–NRPS
(*Cell wall synthesis inhibitor*)

	R$_1$	R$_2$	R$_3$	R$_4$
A	CH$_3$	H	OCH$_3$	CH$_3$
B	C$_2$H$_5$	H	OCH$_3$	CH$_3$
C	CH$_3$	CH$_3$	OCH$_3$	CH$_3$
D	C$_2$H$_5$	CH$_3$	OCH$_3$	CH$_3$
E	CH$_3$	H	H	CH$_3$
F	CH$_3$	H	OCH$_3$	CH$_2$OH
G	C$_2$H$_5$	H	OCH$_3$	CH$_2$OH
H	CH$_3$	H	OCH$_3$	H
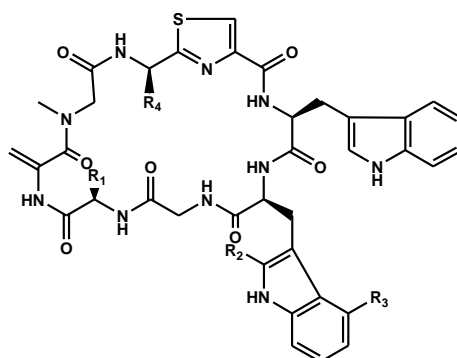
Argyrins-NRPS
(*Immunosuppressive, Antitumour*)
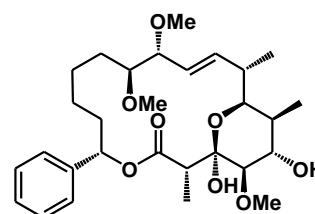
Soraphen-PKS
(*Antifungal*)

R=Ph;   R=CH(CH$_3$)$_2$
YM-47141 & **YM-47142**
(*HLE inhibitors*)
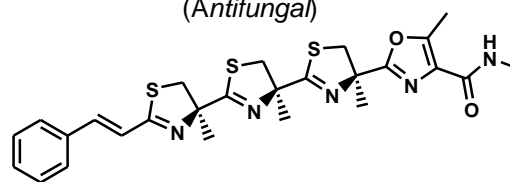
Thiangazole-PKS
(*Anti-HIV*)
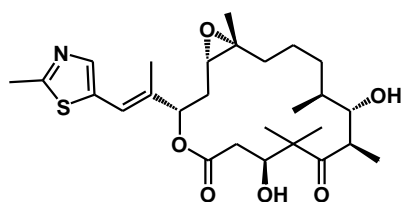
Epothilone B-PKS/NRPS
(*Anticancer-tubulin*)
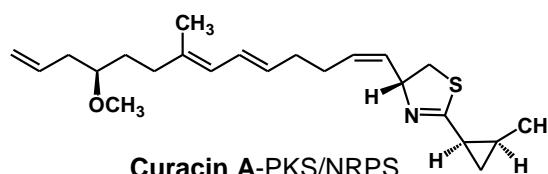
Curacin A-PKS/NRPS
(*Anticancer-tubulin*)

3

Figure 1-1: Compounds from gliding bacteria with diverse activities and modes of action. Compounds inside the box are antibacterial compounds. Biological activity and mode of action are represented in italics. Only compounds, for which the biosynthetic gene clusters are known are depicted. PKS-Polyketide synthase, NRPS-Nonribosomal peptide synthetase, PKS/NRPS-hybrids; *HLE*-Human leukocyte elastase.

1.2. Biosynthetic machineries producing nonribosomal peptides and polyketides

Many secondary metabolites isolated from terrestrial and marine microorganisms are nonribosomal peptides and polyketides. These molecules are constructed by enzymes called megasynthases, i.e nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS), respectively. Both of these machineries are organized in modules, with each module comprising catalytic centres, i.e. domain, to perform the respective function in order to assemble the peptide or polyketide. These domains function in a sequential manner starting up by selecting the specific substrates and subsequently loading them to the enzymatic assembly line via a thioester linkage. The loaded monomers in each module are then connected successively to form the final peptide or polyketide, which may, however be further modified by tailoring enzymes. The last module contains an additional domain called thioesterase (TE) for releasing the assembled product from the megasynthase (Fischbach and Walsh, 2006; Walsh, 2004).

To commence with the operation of megasynthases, the carrier protein (CP) domain that carries the substrate throughout the enzymatic assembly has to be first activated. The priming of the CPs, i.e peptidyl carrier protein (PCP) in NRPS and acyl carrier protein (ACP) in PKS from their inactive apo to their holo form is catalyzed by a phosphopantetheinyl transferase (PPTase) and acyl carrier protein synthase (AcpS), respectively. PPTases and AcpSs are distinct enzymes that transfer the 4'-phosphopantetheine (PPant) group from coenzyme A (CoA) to the hydroxyl group of conserved serine residue in CPs leaving a terminal thiol group (Figure 1-2) (Dall'Aglio et al., 2011; Fischbach and Walsh, 2006). This thiol group would then covalently bind to the substrate via a thioester linkage (Figure 1-3-3) (Walsh et al., 1997). Due to the promiscuity of PPTases and AcpSs, other CoA analogs including acyl-CoAs that are ubiquitous in bacterial cells would act as their substrates rendering the

megasynthases inactive. In that scenario, a TE II domain hydrolyzes the acyl group thereby restoring the megasynthases for the activation by CoA to proceed further with biosynthesis (Heathcote et al., 2001; Schwarzer et al., 2002; Yeh et al., 2004).
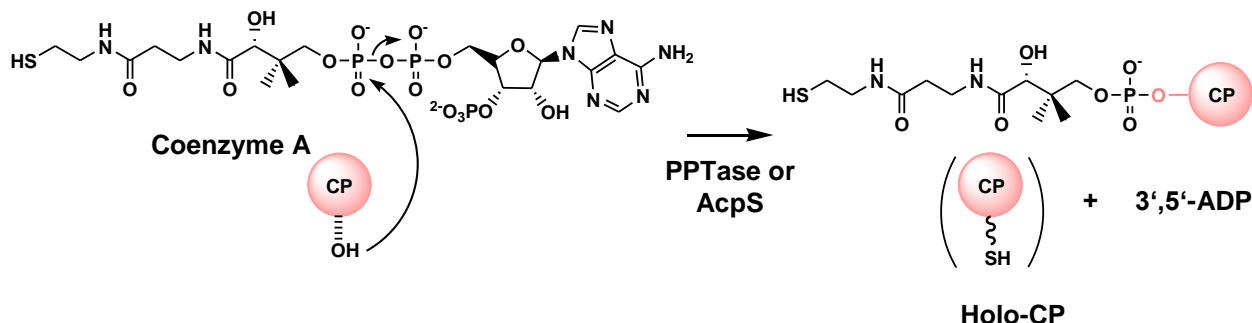
Figure 1-2: Enzymatic priming of carrier proteins (CP) to the active holo form. PPTase-phosphopantetheinyltransferase; AcpS- acyl-CoA protein synthase. The concise representation of holo-CP is in parenthesis are often used in this report. This figure is based on Fischbach and Wash, 2006; Schwarzer et al., 2003.

1.3 Enzymology of NRPS

Nonribosomal peptides synthesized by NRPS modular assembly lines vary from ribosomal peptides primarily in their building blocks and structural features. The building blocks of nonribosomal peptides include non-proteinogenic amino acids in addition to the 20 proteinogenic amino acids. As indicated by their biosynthetic function, an NRPS enzyme contains an initiation, one or more elongation and a single termination module (Finking and Marahiel, 2004). Each of these modules includes an adenylation (A) domain for specific recognition and adenylation of amino acids through the hydrolysis of ATP (Figure 1-3-1). This adenylated intermediate is then loaded on to the thiol moiety of the PPant prosthetic group of the PCP domain, which is located downstream to the A domain of the same module (Figure 1-3-2). Consecutively, the condensation (C) domain of the next module mediates C-N bond formation by Claisen condensation reaction between the upstream aminoacyl(peptidyl)-*S*-PCP and downstream aminoacyl-*S*-PCP. The resulting peptidyl-*S*-PCP intermediate is now ready for the next elongation step (Figure 1-3-3). After the last elongation step the release of the full length peptide from the enzyme is catalyzed by the TE domain (Figure 1-3-4) (Schwarzer et al., 2001). Altogether, an initiation module constitutes A and PCP domains, whereas elongation modules contain C, A and PCP domains. To terminate the biosynthesis, the last elongation

module, should contain a TE domain (termination module) (Fischbach and walsh, 2006).

(1) Activation (adenylation) of the substrates
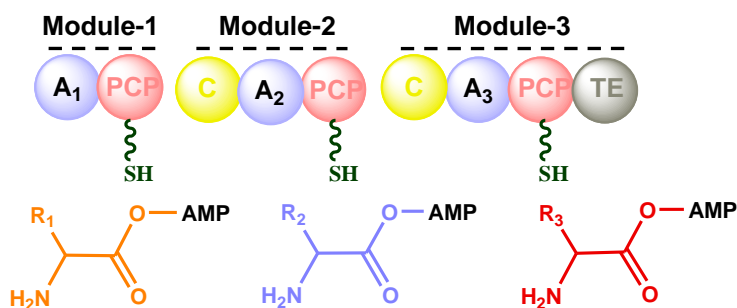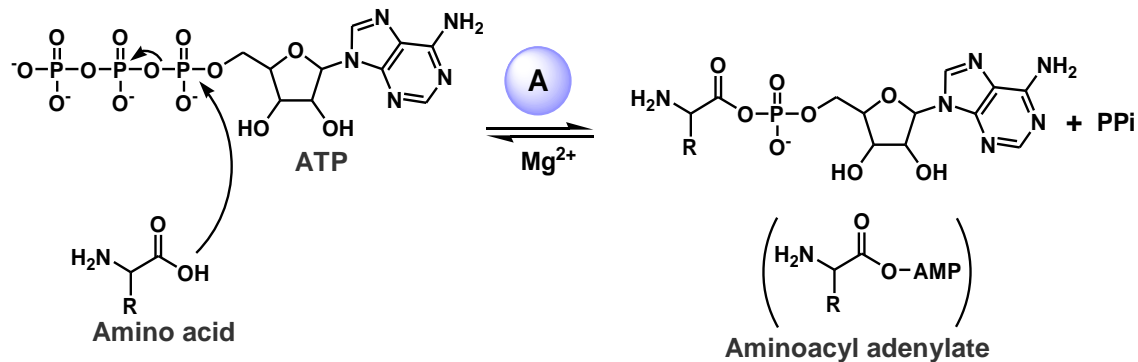
(2) Loading of the activated substrates
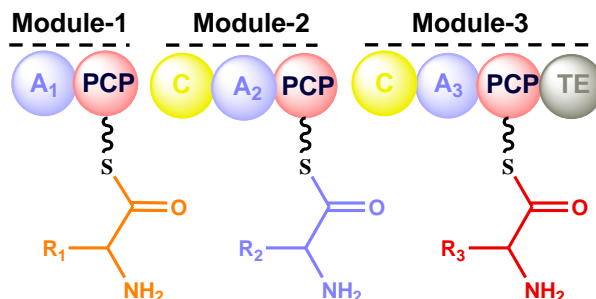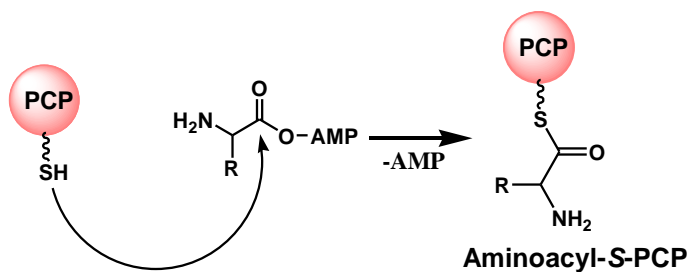
(3) Extension of the chain
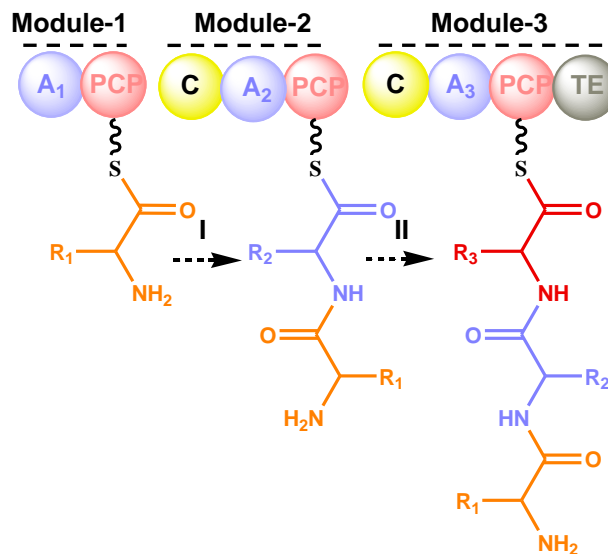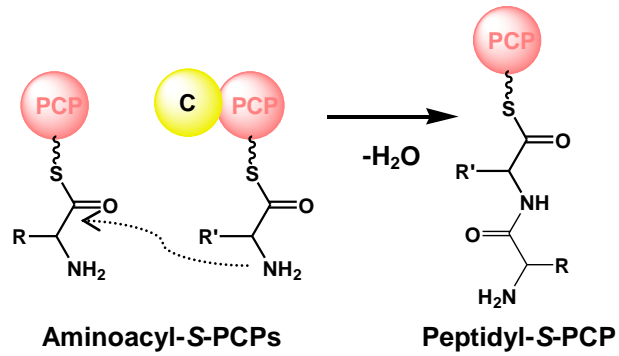
(4) Release of the peptide chain
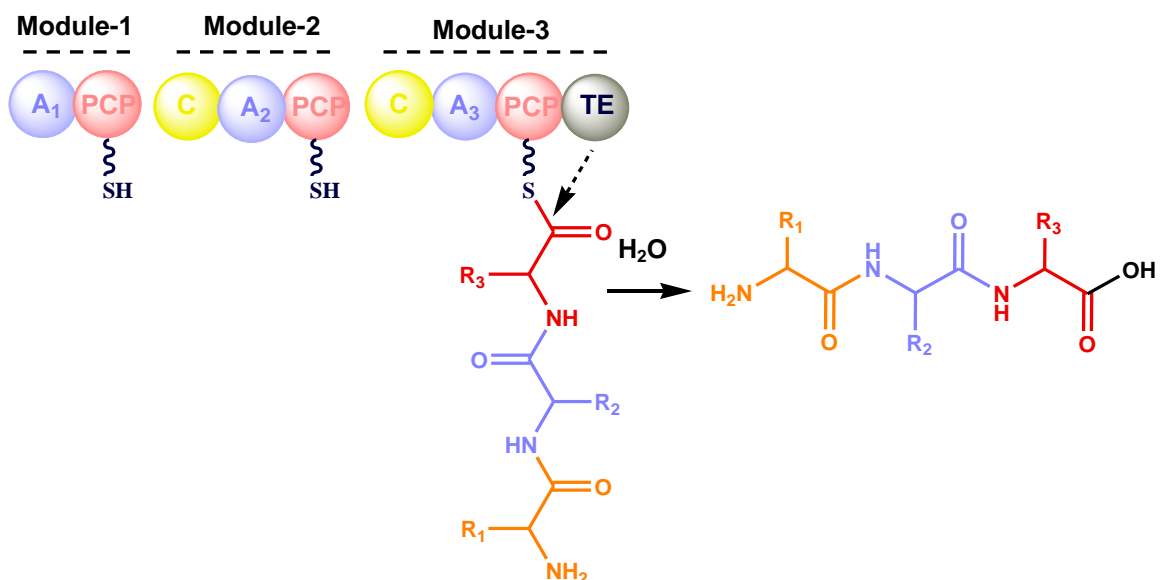
Figure 1-3: Basic steps (1-4) in the biosynthesis of nonribosomal peptides by an activated holo-NRPS. A, adenylation; PCP, peptidyl carrier protein; C, condensation; TE, thioesterase.

R, amino acid side chain; ATP, adenosine triphosphate; Module-1 is the initiation module, module-2 is the elongation module and module-3 is the termination module. Domains involved in each step are written in black. The order of condensation reactions are represented in roman numbers. This figure is adapted from Fischbach and Wash, 2006; Schwarzer et al., 2003.

In addition to the minimal domains, i.e C, A, PCP and TE, NRPS might also contain *cis* acting accessory domains responsible for methylation, epimerization and cyclization providing immense structural diversity to nonribosomal peptides (Samel et al., 2008; Walsh et al., 2001). Depending on the nature of the methytransferase (MT) domain, the aminoacyl-*S*-PCP is either *N*- or *C*-methylated using the methyl group from the cofactor *S*-adenosylmethionine (SAM). Usually *N*-MT is located next to the A domain (Figure 1-4), whereas *C*-MT is located next to the PCP domain (Marahiel, 2009; Schwarzer et al., 2003.).
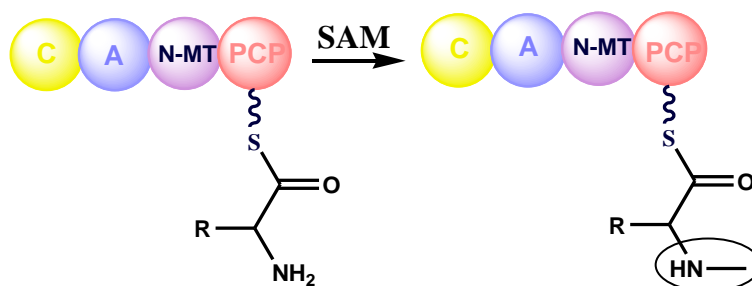
Figure 1-4: Reaction catalyzed by N-methyl transferase (N-MT). SAM-*S*, adenosylmethionine; C, condensation; A, adenylation; PCP, peptidyl carrier protein; R, amino acid side chain. The N-MT domain, which is involved in methylation is written in black. This figure is adapted from Schwarzer et al., 2003.

D-amino acids are prominent in nonribosomal peptides. In cyclosporine and microcystin synthetase, A domains selectively adenylate D-amino acids (Weber and Leitner, 1994). But in most of the NRPS, L-aminoacyl-*S*-PCP has to be epimerized by the action of specialized epimerization (E) domains. E domains that are located next to the PCP domain predominantly mediate the conversion of PCP-bound amino acid from their L-configuration to the D-isomer (Figure 1-5). In certain NRPS like arthrofactin synthetase, the C domain itself performs both the condensation and epimerization reactions (Balibar et al., 2005). This type of dual C domain can be readily distinguished from the conventional C domains based on their sequences.
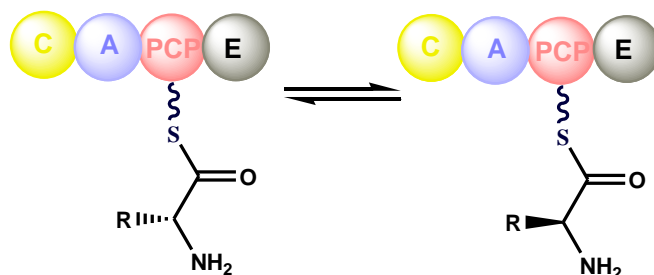
Figure 1-5: Epimerization of PCP bound amino acid by epimerization (E) domain. C-condensation; A-adenylation; PCP-peptidyl carrier protein; R- amino acid side chain. The E domain which is involved in epimerization is written in black.

In addition to the above mentioned domains, another domain known as heterocyclization (HC) domain is of importance. Apart from the peptide bond formation this HC domain catalyzes the nucleophilic attack of a thiol group in cysteine or a hydroxyl group in serine and threonine with an upstream carbonyl group to form a cyclic adduct, followed by dehydration resulting in thiazoline or oxazoline rings, respectively (Schwarzer et al., 2003). This hydrolytically labile dihydroheterocyclic rings can be further oxidized by oxidation (Ox) domain to stable heterocycles, i.e. thiazole and oxazole rings (Figure 1-6) (Schwarzer et al., 2003).
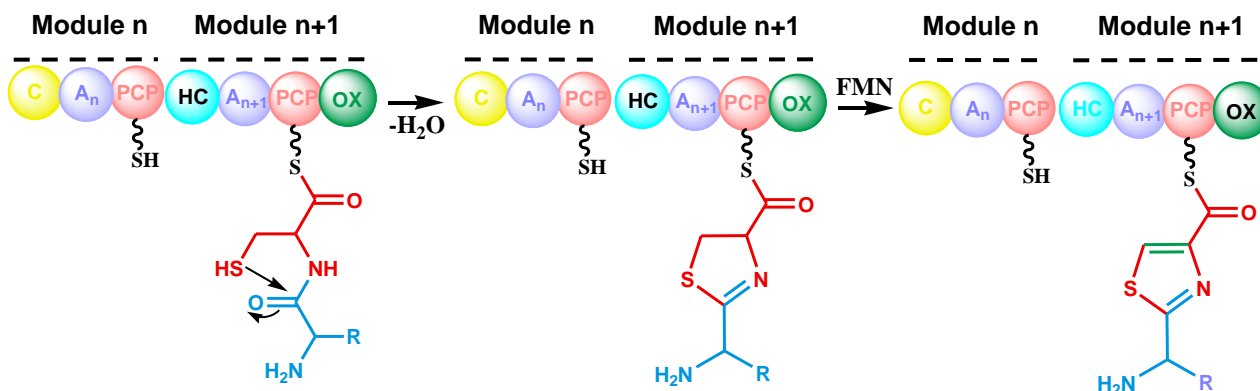
Figure 1-6: Heterocycle formation by susbsequent action of cyclization (HC) and oxidation (OX) domains. C, condensation; A, adenylation; PCP, peptidyl carrier protein; R, amino acid side chain; FMN, flavin mononucleotide. This figure is adapted from Fischbach and Wash, 2006; Schwarzer et al., 2003.

1.4 Enzymology of PKS

PKS function almost in the same way as NRPS, but they obviously differ in their substrates and domains (Walsh, 2007). In PKS systems, the acyltransferase (AT)

domain selects a carboxy acid as substrate (Figure 1-7-1) and loads it on to the acyl carrier protein (ACP) domain (Figure 1-7-2). The ketosynthase (KS) domain mediates the C-C bond formation by decarboxylative condensation between the upstream acyl-*S*-ACP and the downstream ACP bound carboxy acid (Figure 1-7-3). The KS domain mediated elongation proceeds until it reaches the TE domain, which releases the product from the enzyme (Figure 1-7-4) (Hertweck, 2009; Staunton and Weissman 2001).

Unlike the A domains in NRPS, the AT domain selects acyl-CoAs directly from the cells for polyketide biosynthesis. The AT domain in the loading module always prefer monocarboxylic acids such as acetyl-CoA and propionyl-CoA as their substrates. Other unusual substrates include aromatic and branched monocarboxylic acids. But in certain loading AT domains, dicarboxylic acids such as malonyl-CoA can also be selected. In such cases, the KS variant KS^Q decarboxylates malonyl-*S*-ACP yielding acetyl-*S*-ACP. A KS^Q domain which has glutamine instead of the active site cysteine cannot perform the condensation reaction typical for a KS domain. Therefore, KS^Q is needed only in an initiating module that loads dicarboxylic acids (Ikeda et al., 2001; Ligon et al., 2002; Silakowski et al., 1999; Silakowski et al.., 2001). The AT domains of the extending modules always prefer dicarboxylic acids such as malonyl-CoA, methyl malonyl-CoA and ethyl malonyl-CoA as needed for the polyketide backbone. Some AT domains prefer ACP-linked extender units like aminomalonyl-*S*-ACP, hydroxymalonyl-*S*-ACP and its derivative methoxymalonyl-*S*-ACP (Chan et al., 2006) instead. In this case, enzymes required for the biosynthesis of this ACP-bound unusual intermediates is often found encoded within the gene cluster.

(1) Recognition and binding of the substrates by AT domains

Acyl-CoA   Acyl-*O*-AT
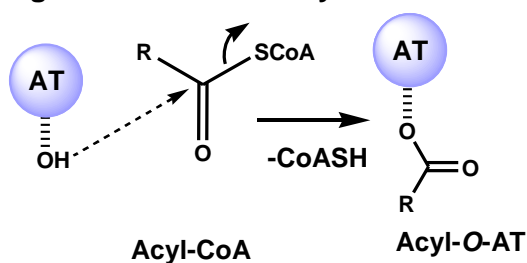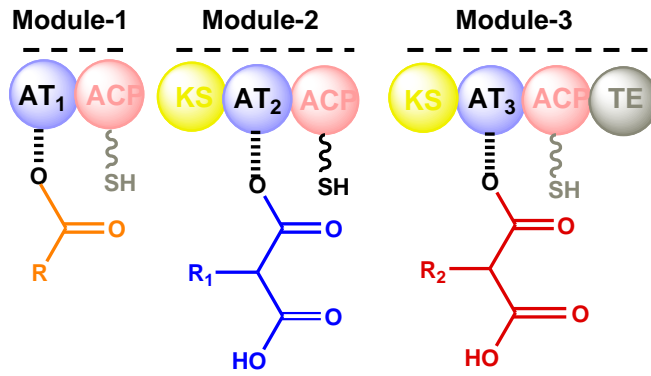
(2) Substrate loading on to the ACP domain

Acyl-*O*-AT   Acyl-*S*-ACP

(3) Chain extension

β-Ketoacyl-*S*-ACP
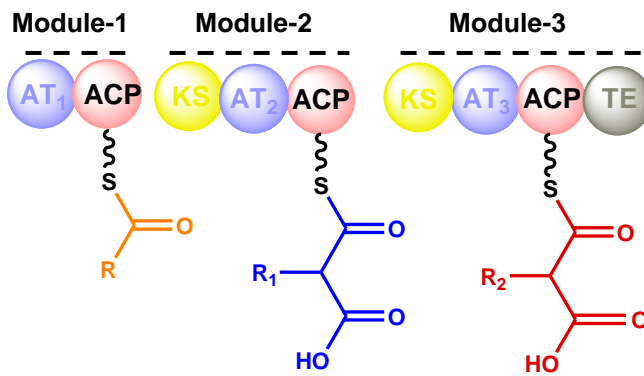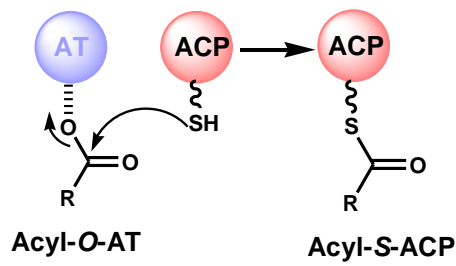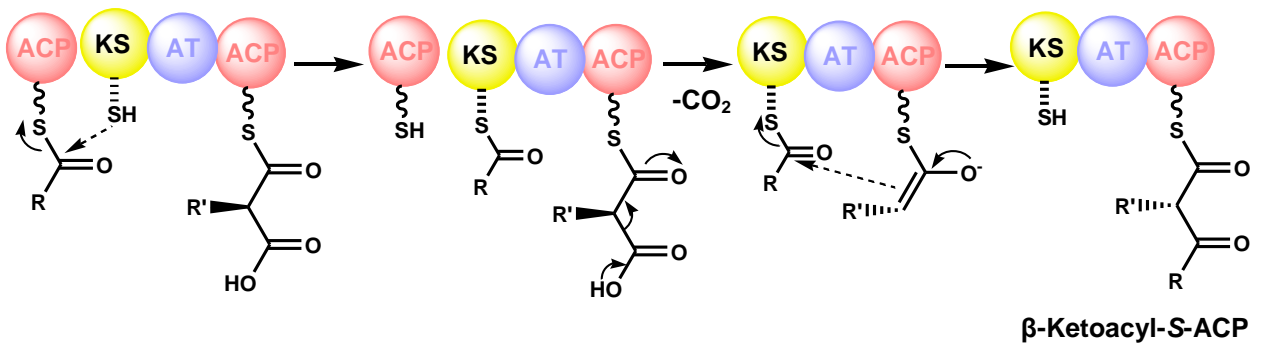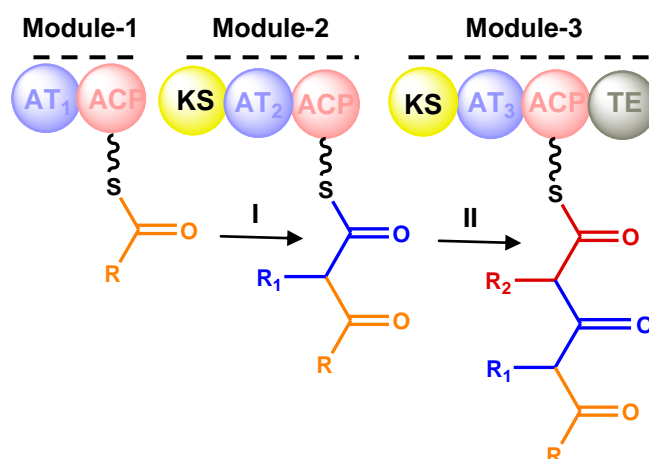
(4) Release of the polyketide chain
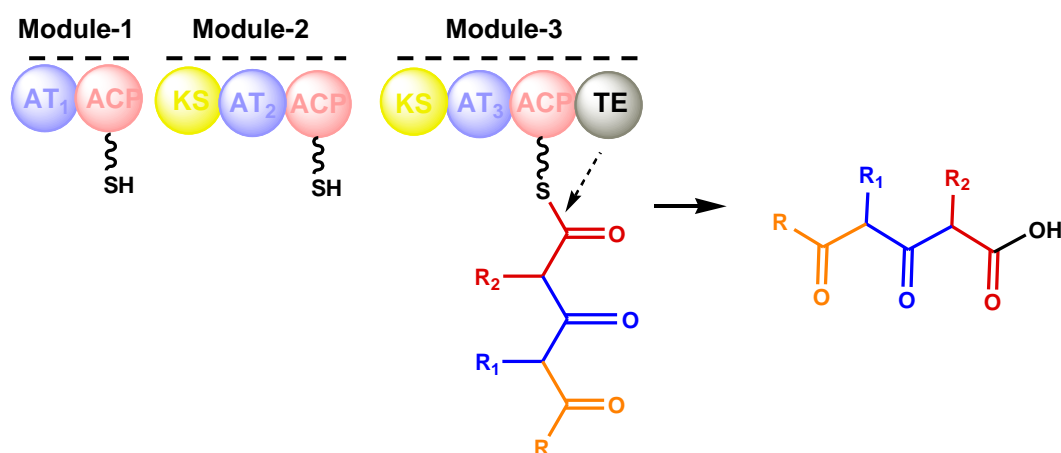
Figure 1-7: Basic steps (1-4) in the biosynthesis of polyketides by an activated holo-PKS. AT-acyl transferase; ACP, acyl carrier protein; KS, ketosynthase; TE, thioesterase; CoA, coenzyme; Module-1 is the initiation module, module-2 is the elongation module and module-3 is the termination module. Domains involved in each step are written in black. The order of extensions is represented in roman numbers. This figure is adapted from Fischbach and Wash, 2006.

Similar to the NRPS, PKS modules may also contain *cis* acting tailoring domains to diversify the polyketide structures (Walsh, 2007). The three catalytic domains ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER) may act sequentially or partially resulting in completely or partially reduced ketide units (Fischbach and Walsh, 2006). The KR domain initially reduces β-ketoacyl-*S*-ACP obtained from the elongation step to β-hydroxyacyl-*S*-ACP. This can be followed by dehydration of β-hydroxyacyl-*S*-ACP by the DH domain to yield a α,β-enoyl-*S*-ACP. This alkene is then saturated completely by the action of an ER domain (Figure 1-8) (Hertweck, 2009).
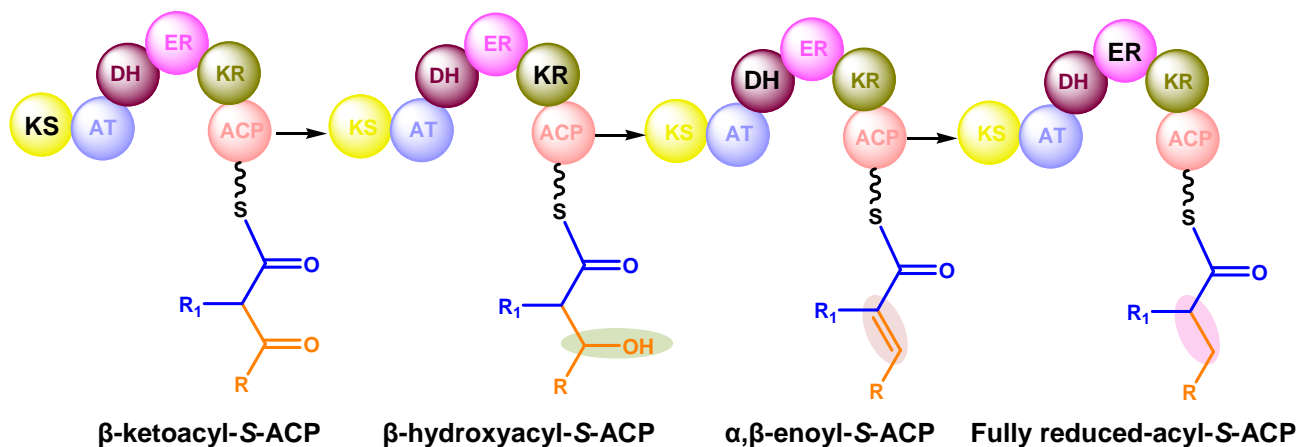
Figure 1-8: β-carbon modifications in polyketides by partial reductions or complete reduction. Subsequent to the formation of β-ketoacyl-S-ACP, ketoreductase (KR) reduces β- keto to a hydroxy group. In the next step, dehydratase (DH) eliminates water from the β-hydroxy function to form the α,β-enoyl-S-ACP, which is then completely reduced by a enoyl reductase (ER) domain. AT, acyl transferase; ACP, acyl carrier protein; KS, ketosynthase; Domains involved in each step are written in black.

Methyl branching at C_α is common and it is either due to the addition of methyl malonyl-CoA as an extender unit or the action of MT as discussed in NRPS. There are also polyketides, like myxovirescin, which are branched at the β position. This β-branching is catalyzed by distinct enzymes called hydroxymethylglutaryl-CoA synthase (HMG-CoA synthase) and enoyl-CoA hydratases (ECH) (Gulder et al., 2011). HMG-CoA synthase mediates the attack of acetyl-S-ACP with β-ketoacyl-S-ACP intermediate. This is followed by the consecutive action of dehydration and decarboxylation by two ECHs resulting in Δ^2-prenyl-S-ACP (Figure 1-9).
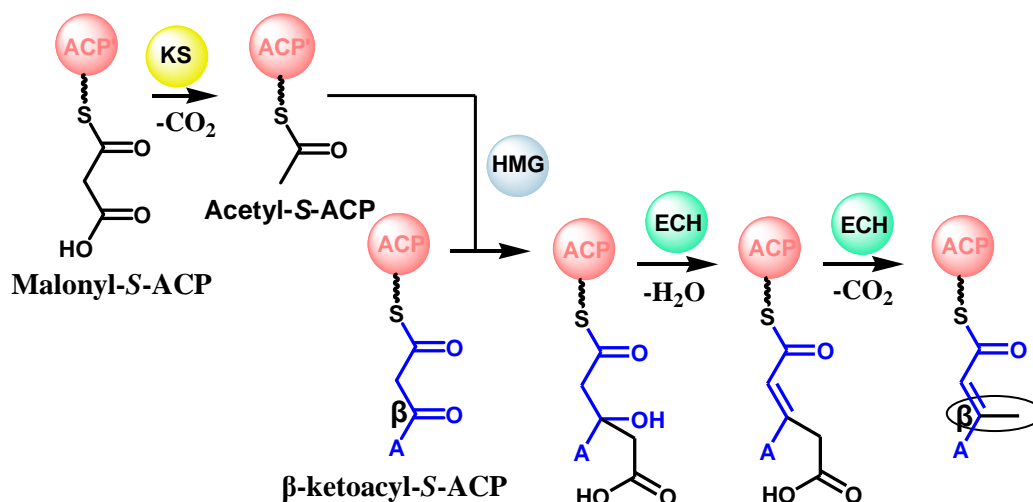
Figure 1-9: Branching at the β carbon (highlighted in oval) catalyzed by the action of hydroxymethyl glutaryl-CoA synthase (HMG) and enoyl-CoA hydratases (ECH). ACP, acyl carrier protein; KS-ketosynthase; A, growing polyketide chain. This figure is adapted from Gulder et al., 2011.

1.5 PKS/NRPS hybrids

Hybrid gene clusters containing both PKS and NRPS modules involved in the biosynthesis of nonribosomal peptide-polyketide hybrid molecules are quite common and sometimes even more predominant than discrete PKS and NRPS gene clusters (Du and Shen, 2001). In such PKS/NRPS hybrid gene clusters, the PKS and NRPS template might be encoded either in the same ORF or in different ORFs. At the NRPS-PKS interface, the KS domain must be able to mediate C-C bond formation between upstream peptidyl-S-PCP and downstream acyl-S-ACP (Figure 1-10A). Similarly at the PKS/NRPS interface, the C domain must be able to catalyze C-N bond formation between upstream β-ketoacyl-S-ACP and downstream aminoacyl-S-PCP (Figure 1-10B). This aberrant shift in the biosynthesis from PKS to NRPS and vice versa requires proper intermodular interaction through linkers for the transfer of the growing chain in the assembly line (Li et al., 2010; Richter et al., 2008). There are exceptions however, e.g. in the biosynthesis of the PKS/NRPS hybrid molecule coranatine, PKS and NRPS modules do not interact. Instead, the peptide and the polyketide scaffold are fused by the action of discrete enzyme (Bender et al., 1998; Rangaswamy et al., 1998a; Rangaswamy et al., 1998b).

(A) NRPS/PKS

(B) PKS/NRPS
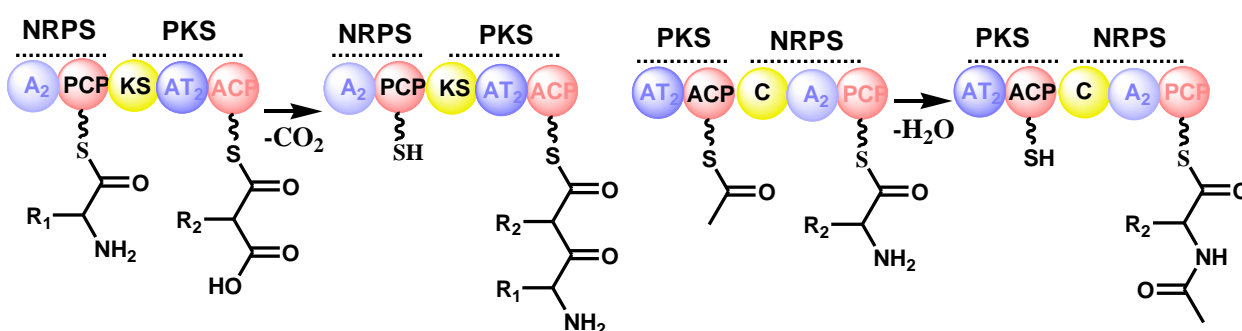
Figure 1-10: NRPS and PKS mixed hybrid systems. A, adenylation; PCP, peptidyl carrier protein; KS, ketosynthase; AT, acyltransferase; ACP, acyl carrier protein.
(A) NRPS/PKS hybrid locus where KS mediates decarboxylative condensation between upstream aminoacyl-S-PCP and the downstream acyl-S-ACP.

(B) PKS/NRPS hybrid locus where C mediates Claisen condensation between upstream acyl-*S*-ACP and downstream aminoacyl-*S*-PCP.

1.6 Nonribosomal peptide and polyketide release mechanisms

The release of a completely assembled chain from the biosynthetic machinery is catalyzed either by the action of TE or reductase (R) domains (Du and Lou, 2010; Kopp and Marahiel, 2007). In the TE mediated chain release mechanism, the full length peptidyl(acyl)-*S*-CP intermediate is transferred to the -OH group of the active site serine residue in TE (Fischerbach and Walsh, 2006). Depending on the nature of the TE the C-terminal carbonyl carbon of peptidyl(acyl)-*O*-TE intermediate is attacked either by an external or internal nucleophile resulting in linear or cyclic structures. Hydrolysis of peptidyl(acyl)-*O*-TE releases the assembled chain as a linear acid (Figure 1-11, Figure 1-12A). In an intramolecular cyclization reaction, a hydroxyl or an amino group of the peptidyl(acyl) chain acts as nucleophile yielding macrolactone or macrolactam rings, respectively (Figure 1-11, Figure 1-12B) (Kopp and Marahiel, 2007). In addition to the above, TE domains observed in enterobactin and gramicidin synthetase catalyze cyclooligomerization of peptidyl(acyl) chains (Figure 1-12C) (Kratzschmar et al., 1989; Rusnak et al., 1991). TE domains that catalyze hydrolytic release or cyclization can be deciphered from the gene sequences, but TEs that catalyze oligomerization cannot be identified on the genetic level. Also the cyclization reaction is highly stereo- and regio-specific, which cannot be determined from the respective gene sequences (Kopp and Marahiel, 2007).
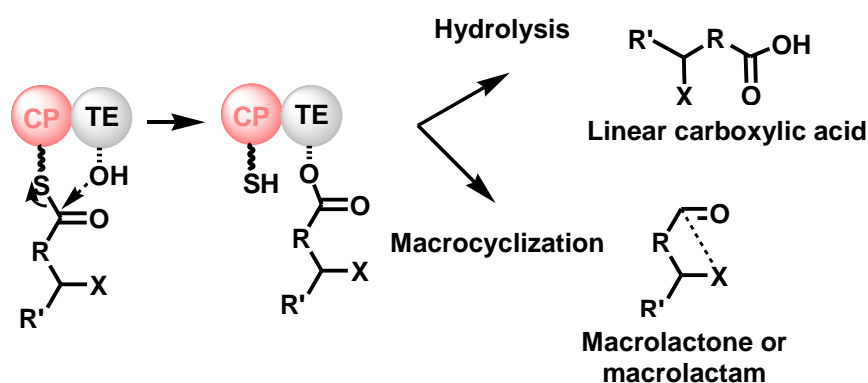
Figure 1-11: General mechanism for the release of a chain from the megasynthase catalyzed by thioesterase (TE). X represents either of the nucleophiles OH or NH_2; R represents acyl(peptidyl) chain; CP, carrier proteins. This figure is adapted from Du and Lou, 2010.

(A) Hydrolysis

Ambruticin J

(B) Macrocyclization

Rifamycin

Epothilone B
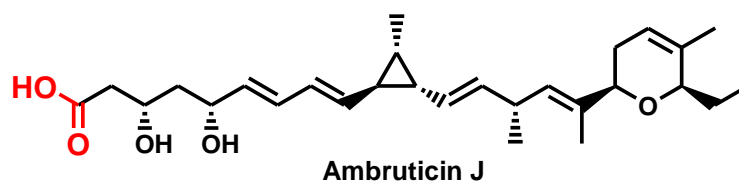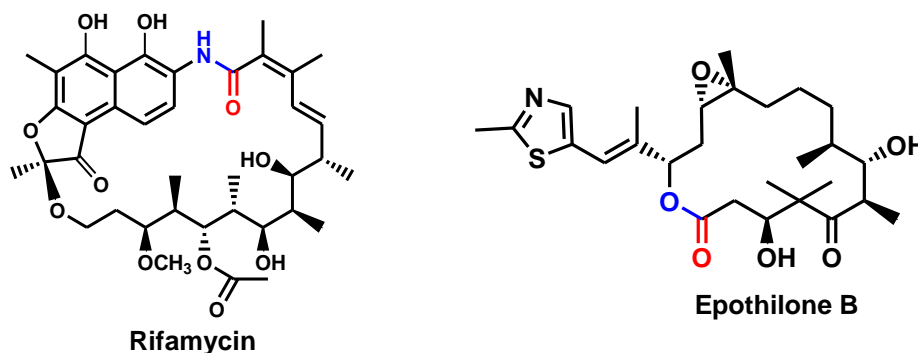
(C) Cyclooligomerization

Enterobactin
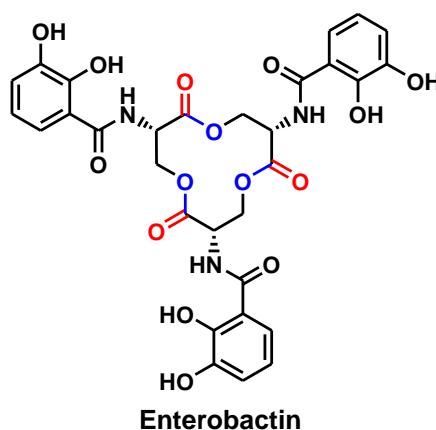
Figure 1-12: Examples of nonribosomal peptides and polyketides dissociated from their megasynthases by the action of TE through hydrolysis (A), macrocyclization (B) and cyclooligomerization (C). Terminal carbonyl thioestered to CP is highlighted in red, whereas the nucleophile that attacks the thioestered carbonyl is highlighted in blue.

In the absence of a TE domain, a R domain may catalyze the reductive release of peptidyl(acyl)-S-CP as a linear aldehyde from the megasynthase (Figure 1-13A) (Kopp and Marahiel, 2007). As observed in myxochelin A, R domains might also act iteratively thereby reducing the aldehyde to an alcohol (Figure 1-13B) (Gaitatzis et al., 2001). Recently, a variant of the R domain called Dieckmann type R domain was identified in PKS/NRPS hybrid gene clusters such as tenellin and equisetin synthetase (Figure 1-13D). Instead of reduction, this variant catalyzes Dieckmann

cyclization (Figure 1-13C) (Sims and Schmidt, 2008). This variant can be distinguished from the R domain by its amino acid sequence and a conserved aspartate residue (D3803) (Liu and Walsh, 2009).

(A) Reductive release

(C) Dieckmann cyclization

(B) Example for reductive release

(D) Example for Dieckmann cyclization

Myxalamide A

Equisetin
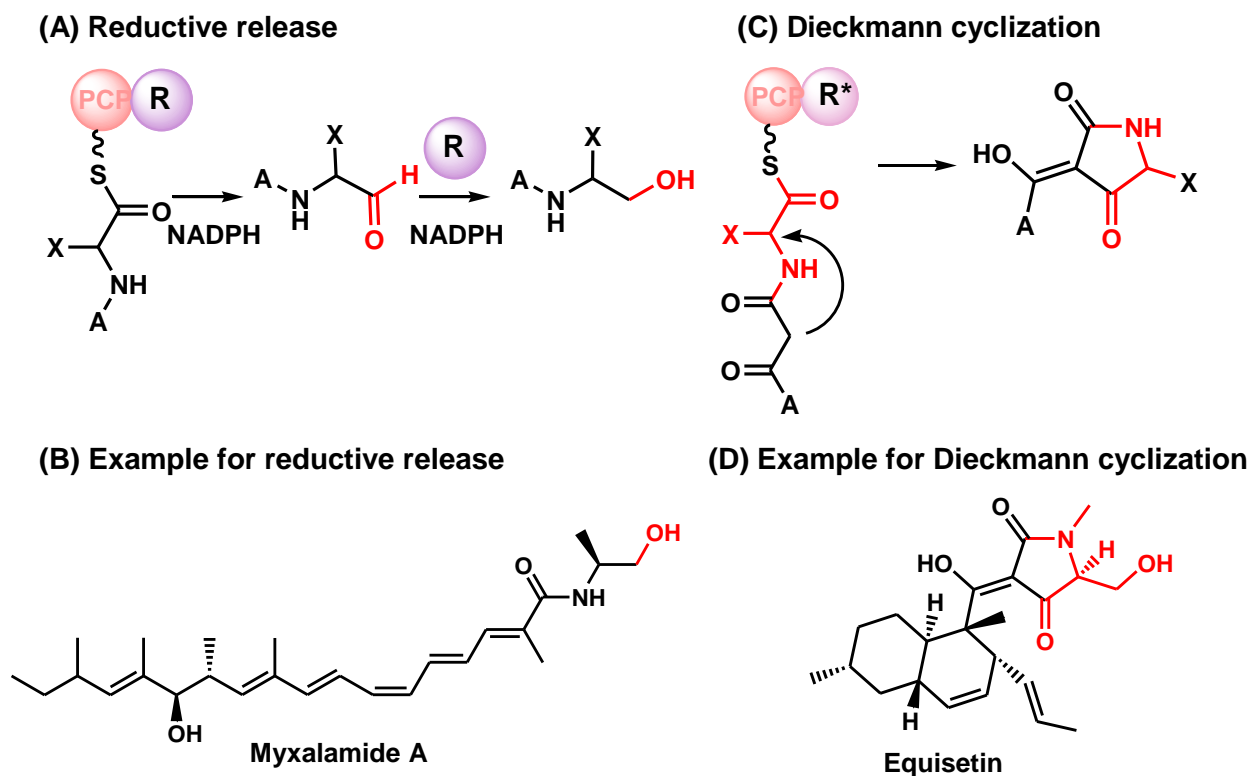
Figure 1-13: Reductase (R) domain release mechanism.

(A) Reductive release of the growing chain as an aldehyde or alcohol by the action of a reductase (R) domain. NADH-nicotinamide dinucleotide; NADPH-nicotinamide dinucleotide phosphate

(B) Reductively released myxalamide A as an alcohol.

(C) Release of the peptide-polyketide hybrid by R domain variant R* via Dieckmann cyclization.

(D) Equisetin dissociated from equisetin synthetase via Dieckmann cyclization.

These figures are adapted from Du and Lou, 2010.

Biosynthetic gene clusters without TE or R domain were reported to mediate the chain release by alternate mechanisms. However, with the identification of more PKS and/or NRPS gene clusters without any specialized domains for release, new chain release mechanisms will continue to emerge (Du and Lou, 2010).

1.7 Post-assembly tailoring reactions

Many of the nonribosomal peptides and polyketides undergo further modifications such as glycosylation, methylation, halogenation and oxidation after their release from the respective megasynthases (Figure 1-14) (Samel et al., 2008; Walsh et al., 2001). These modifications often influence the bioactivity, selectivity and pharmacokinetic properties of the molecules (Kopp et al., 2007; Thibodeaux et al., 2008). For instance, non chlorinated analogue of the antibiotic bahimycin showed drastic reduction in their activity (Bister et al., 2003).The genes encoding the enzymes glycosyltransferase (GT), MT, halogenase (H), monooxygenase (MO) for the modification reactions are often found within the gene cluster (Walsh, 2007). In rare cases like chivosazol synthetase, GT acts in *trans* from a remote location with respect to the gene cluster (Perlova et al., 2006).

Chivosazol A

Jerangolid A
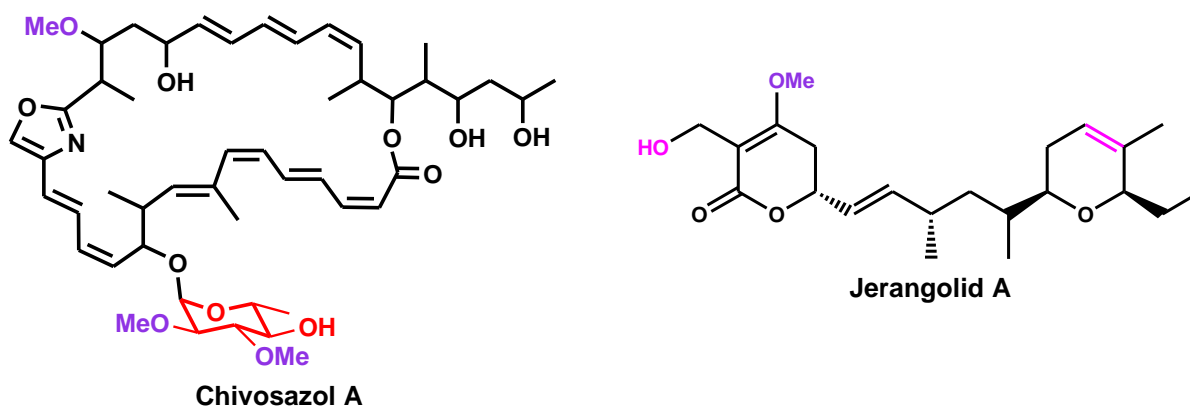
Figure 1-14: Examples for post-assembly modification reactions. The addition of a glycosyl group by the action of a gylcosyltransferase (GT) is highlighted in red, *O*-methylation by *O*-MT is highlighted in violet, double bond and hydroxylation by the action of a monooxygenase is highlighted in pink.

1.8 Primary structure determinants in NRPS and PKS

1.8.1 NRPSs 'A' domains

The A domain in NRPS modules determines the specific amino acid that has to be incorporated into the nonribosomal peptide. This domain activates the amino acid as aminoacyl-AMP for its loading on to the PPant arm of PCP domain. Detailed studies on the 3D structure of a phenylalanyl-AMP bound A domain, i.e PheA of gramicidin synthetase (GrsA) revealed ten residues at positions 235, 236, 239, 278, 299, 301, 322, 330, 331 and 517 that are present in the conserved motifs A3 to A7 and A10

make direct contact to phenylalanine (Figure 1-15) (Conti et al., 1997; Lautru and Challis 2004). These ten residues that are critical for the substrate binding, control the entry of the specific substrate and are called as the 'Stachelhaus code' or 'nonribosomal code'. Sequence alignments of the characterized A domains with GrsA synthetase enabled the extraction of the Stachelhaus code pointing towards the respective substrates (Challis et al., 2000; Stachelhaus et al., 1999). This in turn facilitates the identification of substrates for an uncharacterized A domain from its Stachelhaus code. Domain identification and substrate prediction lays the foundation for predicting a nonribosomal peptide structure from the respective genes.

(A)   **(B)**
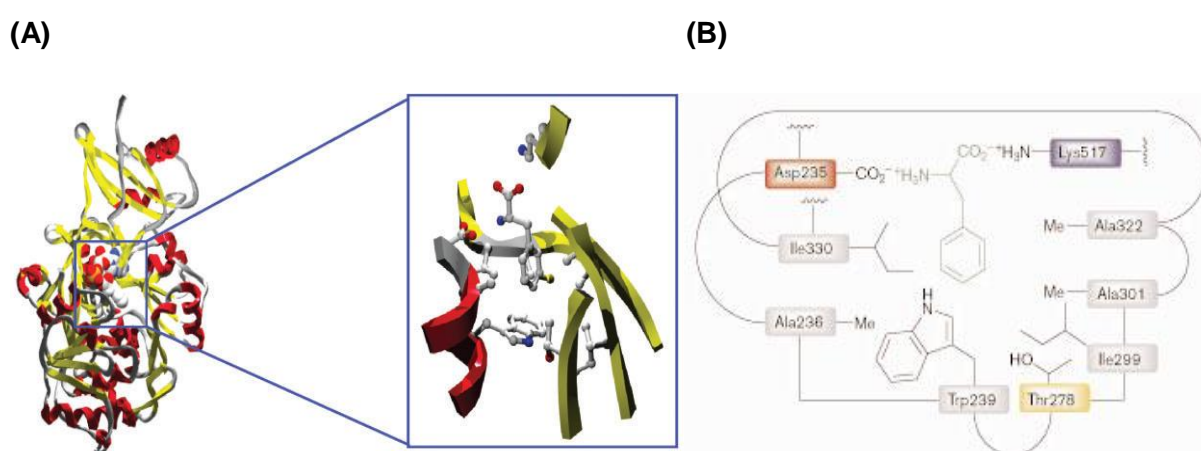
Figure 1-15: Adenylation (A) domain: the structure determinant in NRPS.

(A) Structure of the L-Phe-activating domain (PheA) of the gramicidin S synthetase GrsA, with an enlargement of the L-Phe-binding pocket showing the residues of PheA making contact with L-Phe. The side chains of these residues and the substrate L-Phe are rendered in ball and stick representation. This figure is reproduced from Lautru and Challis, 2004.

(B) Two dimensional representation of L-Phe binding pocket in PheA of gramicidin S synthetase GrsA. Residue colours indicate its nature; red-acidic; blue-basic; yellow-neutral polar; gray-hydrophobic. This figure is reproduced from Challis et al., 2000.

1.8.2 PKSs AT domains

Unlike the A domains, the AT domains of PKS selectively catalyze the transfer of acyl groups from acyl-CoA to the PPant arms attached to the active site serine of the ACP domains (Hertweck, 2009). Extensive analysis of *E. coli* FAS AT and 6-deoxyerythronolide B synthase (DEBS) AT crystal structures has identified four signature motifs that govern substrate specificity (Smith and Tsai, 2007; Tsai and Ames, 2009). The AT structure includes two domains: a larger core subdomain and a

smaller subdomain. The catalytically active serine (GH**S**XG) in motif-2 lies in the cleft formed between these two subdomains. Motif-1 and -3 which is 30 and 100 residues upstream and downstream to the active site serine lies on both sides of the catalytic cleft (Figure 1-16). Therefore, changes in these motifs affect size and shape of the substrate binding pocket. As per mutational studies in DEBS AT domains the motif-3 is dominant in determining the substrates. Furthermore, α-β-α motif-4, a hypervariable region at the C-terminus has been identified to be important in recognizing the specific substrate. But based on amino acid sequence in this region to date no possible prediction about the substrates was made yet (Smith and Tsai, 2007; Tsai and Ames, 2009). However, consensus sequences in motif-1, -2 and -3 have been categorized between the extender units, malonyl and methylmalonyl-CoA. But a general pattern in identifying AT domains selecting unusual extenders and different starters is not yet fully developed. Furthermore, in AT domains selecting dicarboxylic acids a conserved arginine at position 117 is necessary for interacting with the C3 carboxylate oxygen, whereas a nonpolar amino acid is usually observed at this position in case of monocarboxylic acid substrates (Smith and Tsai, 2007; Tsai and Ames, 2009). In contrast to the NRPS A domains, prediction of especially the unusual substrates from PKS AT domains is not straightforward and cannot be done with accuracy.

(A)   **(B)**

Motif	Malonyl-CoA	Methylmalonyl-CoA
Motif-1	*ZTX$[AT][QE]* *Z*- hydrophilic amino acid *$*- aromatic amino acid	*[RQSED]V[DE]VVQ*
Motif-2 *GHSXG*	*X*-branched hydrophobic amino acid	ATs except malonyl-AT *X*-less bulky residue
Motif-3	*[HTVYGMIL]AFH* *GASH*	*[YVW]ASH*
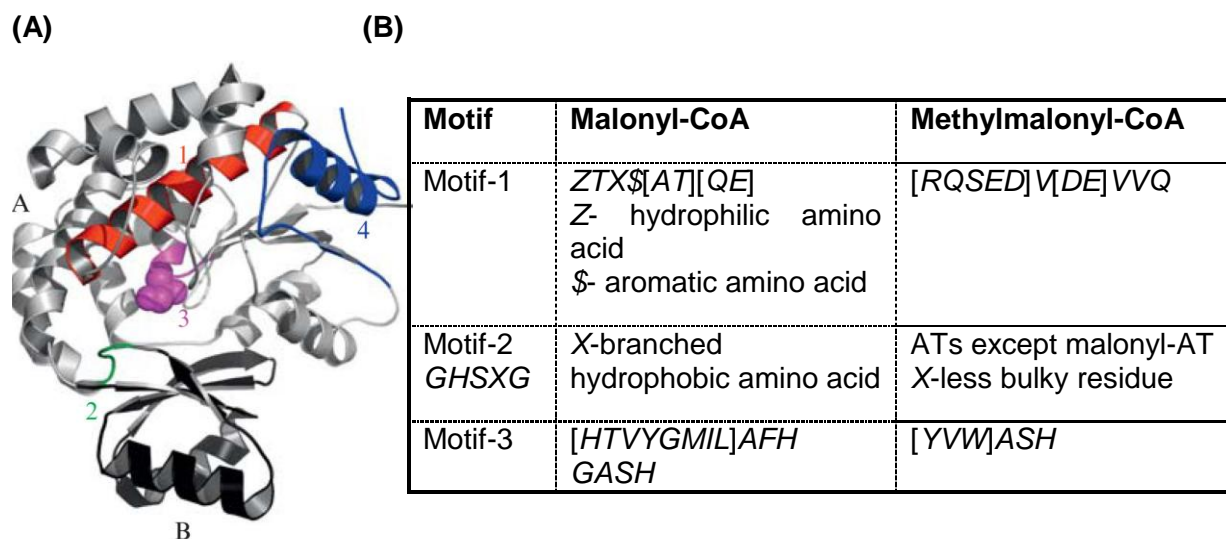
Figure 1-16: Acyl transferase (AT) domain, the structure determinant of PKS.

(A) Overall fold of AT domains (large core subdomain (A) in grey, small subdomain (B) in black), and the four substrate motifs. Motif-1 in red, motif-2 in green, motif-3 with serine in purple, and motif-4 in blue. This figure is reproduced from Tsai and Ames, 2009.

(B) Consensus sequence of motif-1, -2 and -3 in malonyl- and methylmalonyl-specific AT domains.

1.8.3 Stereochemistry of the natural product deduced from NRPS and PKS sequences

NRPS and PKS generate molecules with a vast diversity concerning the structural template and its stereochemistry. Analogous to the prediction of structure backbones, the stereochemistry implemented by NRPS and PKS assembly lines can be predicted to a certain extent.

1.8.3.1 Stereospecificity of C domains in NRPS

In NRPS, D-amino acids are often formed by the action of E or dual C domains. However, the successful incorporation of these D-amino acids into the assembly is done by the stereo-specific C domains (Schwarzer et al., 2003). The C domain has two binding sites; one for the electrophilic donor, i.e upstream aminoacyl(peptidyl)-*S*-PCP and the other for the nucleophilic acceptor, i.e downstream aminoacyl-*S*-PCP, respectively (Figure 1-17) (Finking and Marahiel, 2004). Since both binding sites exhibit strong stereoselectivity (L- or D-amino acid), conserved residues driving the stereospecificty of C domains are observed in these binding sites. Based on this, C domains that are specific for D configured upstream donor and L configured downstream acceptor are classified as ^DC_L domain, whereas domains that are specific for L configured donor and acceptor are termed as ^LC_L domains (Rausch et al., 2007). From around 300 C domain amino acid sequences, the residues that differ among ^LC_L and ^DC_L domains were proposed. Especially a moderately conserved motif LPxDxxRP is usually observed in ^LC_L domain at the N-terminus (Rausch et al., 2007).
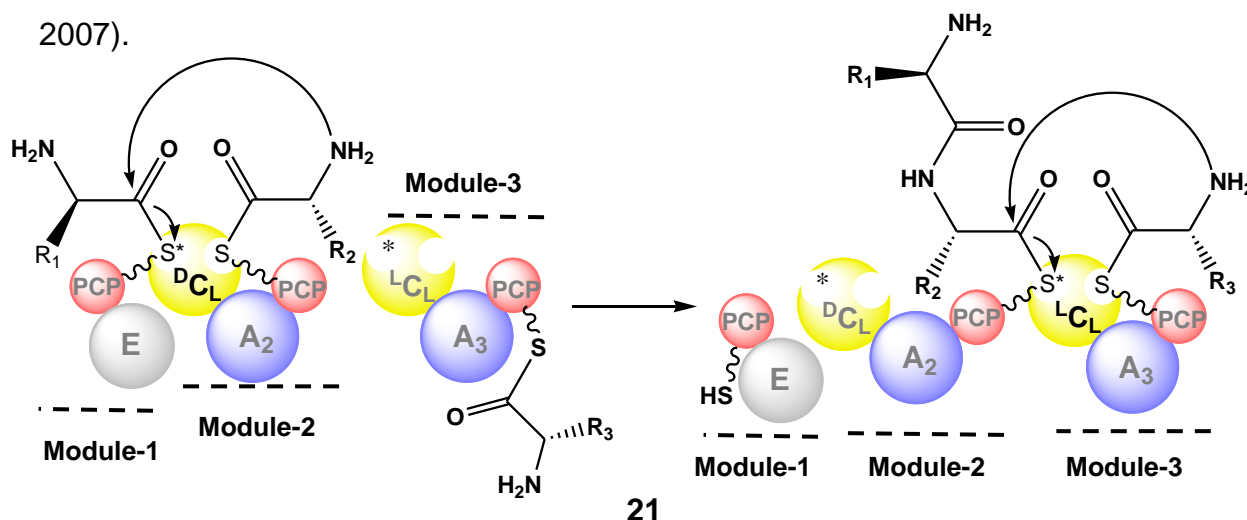
Figure 1-17: General mechanism of condensation (C) domains. The white pockets in the C domains are the acceptor and donor sites where the electrophilic donor and the nucleophilic acceptor, respectively accommodate for the condensation reaction. The acceptor site is marked with an asterisk. The prefix and suffix of the C domain, i.e ^LC_L and ^DC_L indicates the specificity of the acceptor and donor sites, respectively for the L- or D- configured residue. A, adenylation; PCP, peptidyl carrier protein. This figure is adapted from Finking and Marahiel, 2004.

1.8.3.2 Stereospecificity of KR domains in PKS

The significant feature of polyketides is the presence of methyl-, hydroxyl- and double bond stereogenic centres. The KR domain catalyzes the reduction of β-ketoacyl-*S*-ACP with the aid of NADPH to β-hydroxyacyl-*S*-ACP generating either D-hydroxy or L-hydroxy intermediates (Caffrey, 2003). This variation in hydroxyl stereochemistry is due to a difference in the binding orientation of the substrate to the KR active site. This change in orientation is guided by the change in the amino acid sequences. A KR which generates L-β-hydroxacyl-*S*-ACP has a conserved tryptophan residue. This type of KR is termed as A-type KR. In the B-type KR, which generates D-β-hydroxacyl-*S*-ACP, a conserved leucine-aspartate-aspartate (LDD) motif is observed (Kwan and Schulz, 2011; Valenzano et al., 2009) (Figure1-18)

In addition to hydroxyl stereochemistry, KR guides the epimerization of α-methyl substituents of β-ketoacyl-*S*-ACP. Based on tylosin synthase KR crystal structure, conserved residues are observed to participate in the epimerization of α-methyl group in β-ketoacyl-*S*-ACP intermediates (Kwan and Schulz, 2011; Valenzano et al., 2009). Based on the configuration of α-methyl group, A- and B-type KRs are further subdivided into A1, A2, B2 and B2, where 1 refers to D-isomer and 2 refers to L-configuration.

KRs without NADPH binding site are often considered inactive and thereby proposed to be skipped during biosynthesis. But in narbonolide synthase, one of the KR domains that lacks the NADPH binding motif, but with the catalytically active triad lysine-serine-tyrosine was observed to catalyze the epimerization of an α-methyl group (Figure-1.8A) (Zheng and Keatinge-Clay, 2011). This type of KRs without NADPH binding motif but able to perform the epimerization of α-methyl are termed as C2-type.The KRs that are inactive lacking in both NADPH binding site and catalytic

active residues are referred as C1-type KR (Figure-1.8B) (Zheng and Keatinge-Clay, 2011).

(A)
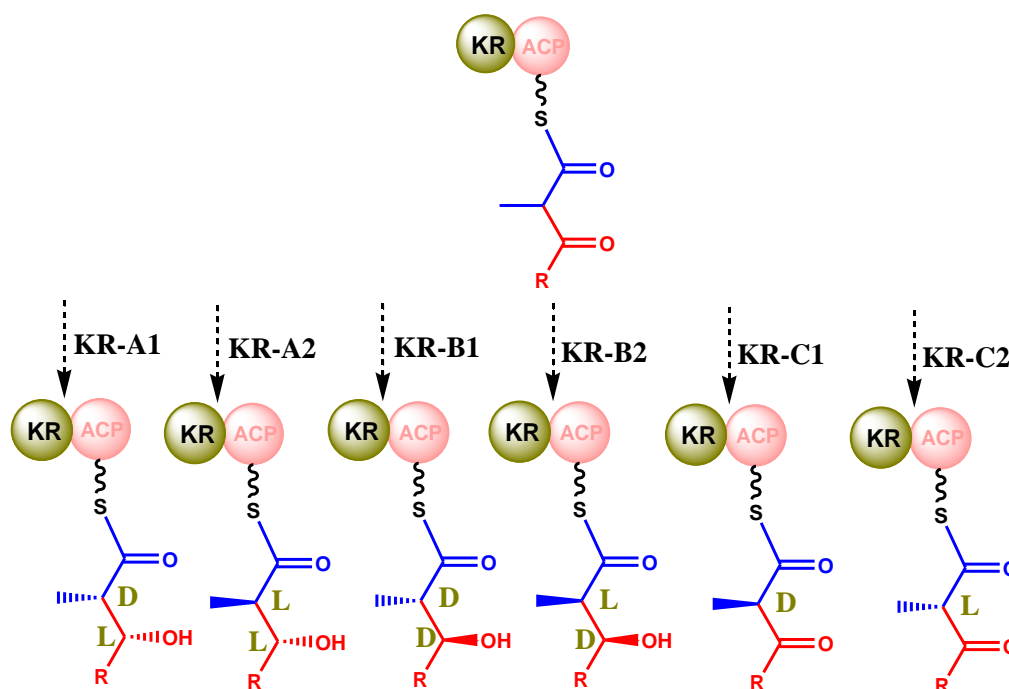
(B)

KR-type	Conserved residues
KR-A1	*Trp141*
KR-A2	*Trp141*, His146
KR-B1	*Leu93*, *Asp94*, *Asp95*,Pro144, Asn 148
KR-B2	*Leu93*, *Asp94*, *Asp95*,Pro144, Asn 148,Pro151
KR-C1	None (Catalytic active residues are also absent)
KR-C2	None but possess the catalytic active triad for KRs

Figure 1-18: Stereochemistry of ketoreductase (KR) domain.

(A) Stereochemistries concerning α-methyl and β-hydroxy group imposed by KR domains.

(B) Based on the conserved residues present, KRs catalyze the epimerization of ACP bound β-ketoacyl intermediates. Residues that differentiate A- and B-type KRs are highlighted in bold italics. This table is based on Zhen and Keatinge-Clay, 2011.

1.8.3.3 Stereospecificity of ER domains in PKS

The stereochemistry of an α-methyl moiety is also reflected in the amino acid sequences of ER domains. Swapping the domains combined with site directed mutagenesis studies showed conserved residues necessary for the epimerization of an α-methyl group (Table 1-2) (Kwan and Leadlay, 2010). Similar to KR, ERs with L- and D-configured α-methyl are termed as ER1 and ER2, respectively (Zheng and Keatinge-Clay, 2011). Thus, the stereochemistry of polyketides especially α-methyl groups is suggested to be determined by the combined action of KR and ER domains.

ER-type	Conserved residues
ER1	Leu46, Ile47, Tyr52, Pro53
ER2	Val46, Val47, Val52

Table 1-2: Stereochemistries concerning the α-methyl groups imposed by an ER domain. This table is based on Zhen and Keatinge-Clay, 2011.

The identification of domains, their substrates and the resulting stereochemistry would enable the prediction of the encoded putative structures. This would indeed also facilitate the identification of the respective biosynthetic genes from the structure investigated. The stereochemistry of chivosazol and etnangien was first *in silico* predicted, and later it was found to be indeed in agreement with the chemically elucidated structures (Menche et al., 2008; Perlova et al., 2006).

1.9 Challenges in genes to structure prediction

The structure prediction from genes is done by considering the 'co-linearity' between the number of modules and the number of building blocks. But the megasynthases are not always constructed in a straightforward way and vary in their organization and functionality (Haynes and Kallis, 2007; Mootz et al., 2002; Schwarzer et al, 2003).

The classical type NRPS displays a co-linearity between structure and modular assembly and is termed linear or type-A NRPS. This type-A NRPS is more common and the product normally predictable from the genes. In enterobactin synthetase the TE domain catalyzes the iterative action of modules for the assembly of enterobactin,

i.e. oligomerization (Kratzschamer et al., 1989). These types of NRPSs are termed as iterative or type-B NRPS. Since the iterative mechanism guided by the TE domain is not yet predictable from the primary genetic structure, type-B NRPS cannot be identified from genes. Another, however rare type of NRPS exists, where domains are organized in a manner different from that of the linear or type-A NRPS. These are the non-linear or type-C NRPS (Haynes and Kallis, 2007; Mootz et al., 2002; Schwarzer et al, 2003). In type-C NRPS, the unusual arrangement of domains is readily identifiable from the genes, but the order of the reactions is not easy to predict.

Similar to NRPS, 'non-canonical' PKSs were found and they are far more complicated than NRPSs (Hertweck 2009; Li and Müller, 2009). Type-I PKS with large multi-modular assemblies are subdivided into two types: Linear type-I PKS where the catalytic domains are used only once in a co-linear fashion, and iterative type-I PKS where the catalytic domains are used more than once (Fischbach and Walsh, 2006; Schumann and Hertweck, 2006). Since, the iterative action cannot be predicted form gene sequences, only very preliminary product structures of linear type-I PKS can be determined. In Type-II PKS, the required catalytic domains are encoded as discrete proteins. These proteins must interact to form a functional PKS enzyme complex, where the catalytic active sites are used either once or iteratively (Brachmann et al., 2007; Sandmann et al., 2007). As the order of extensions is hard to predict from the sequences, type II PKS can only be identified on the genetic level but not the structure of the product. Type-III PKS are small dimeric proteins that catalyze iterative condensations of malonyl-CoA and are independent of an ACP domain for the assembly (Moore and Hopke, 2001). Again the iterative action cannot be predicted from the sequence.

1.10 The 'era' of genome mining

The search for new compounds from natural resources employed for a long time a bioactivity or chemistry-guided approach, whereby bioassays or spectroscopic methods accompanied the isolation process. Once the structure of the pure metabolite was elucidated, its biosynthesis was delineated on the biochemical and often also on the genetic level. These investigations gave insight into the organisation of biosynthetic machineries like those for nonribosomal peptides and

polyketides. Comparative analyses of NRPS and PKS genes and enzymes of known compounds facilitated the identification of consensus motifs indicative for the enzymatic activity of the encoded protein, e.g. based on the so called Stachelhaus code it is possible to predict the substrate likely to be selected by the A domain of an NRPS module (Challis et al., 2000; Stachelhaus et al., 1999).

In the 1990's bacterial genome sequencing commenced, which gave, apart from a deeper insight into bacterial metabolism and possible antibiotic drug targets, a detailed understanding of natural products biosynthesis. These genomes together with the knowledge on biosynthetic genes added a new dimension to natural product research. The analysis of the genome of *Streptomyces coelicholor* A3(2) revealed the presence of biosynthetic genes over-representing the number of isolated compounds. In total 22 biosynthetic gene clusters were identified, however only four products namely prodiginines, actinorhodin, Calcium-dependent antibiotic (CDA) lipopeptide and a pigment had been isolated before (Lautru et al., 2005). An NRPS system with genes required for the synthesis of a peptidic siderophore was the first structure predicted from this genome sequence (Challis and Ravel, 2000). The cultivation of *S. coelicholor* A3(2) in an iron depletion medium resulted in the isolation of this siderophore, which is a tetrapeptide named coelichelin(Lautru et al., 2005). This finding resulting from genome mining has accelerated the search for new biosynthetic pathways in other sequenced genomes, and resulted in new biosynthetic pathways not only in *Streptomyces*, but also in other taxa including cyanobacteria, myxobacteria, bacilli (Challis, 2008; Corre and Challis, 2009; Lanen and Shen, 2006; Winter et al., 2011) New biosynthetic gene clusters were even found in well studied groups of fungi and plants (Bergmann et al, 2007; Bok et al., 2006; Fazio et al., 2004).

1.11 Exploiting the power of genomes

A prerequisite for searching genomes for biosynthetic genes are powerful genome mining tools, which in many cases also allow to predict the chemical nature of the product of an identified biosynthetic pathway (Corre and Chalis, 2007). There are several bioinformatic tools developed for the identification of modular genes from gene sequences including CLUSEAN (Weber et al., 2009), PKS/NRPS analysis (Bachmann and Ravel, 2005) and NRPS-PKS database (Yadav et al., 2003). Multiple

sequence alignment of the PKS AT domains enables the prediction of their respective substrates to a certain extent. Unlike AT domains, special programs like NRPSpredictor (Rausch et al., 2005) and NRPSpredictor2 (Rottig et al., 2011) has been developed to identify the A domain substrates. Using these tools, many gene clusters are identified from the genomes, but the corresponding metabolites for many of these gene clusters, then called cryptic gene clusters, are not yet isolated from the producer. Partly they were overlooked before, due to their presence in very low amounts, thereby escaping the traditional screening methods, or expression of the biosynthetic genes does not take place under laboratory conditions.

There are different strategies, some of which are complementary, for the detection and isolation of the products of cryptic gene clusters (Figure 1-19). A more traditional approach includes the variation of culture conditions (temperature, media, cultivation vessel etc.) in order to induce the expression of the desired biosynthetic pathway [called "One Strain Many Compounds Approach, OSMAC"] (Christian et al., 2005; Cueto et al., 2001). Compounds, whose proposed structure suggests special physiochemical properties, i.e. UV maxima, NMR resonance signals, may be detected by making use of these features (Zazopoulos et al., 2003). Feeding of one of the predicted building blocks, e.g. isotopically labeled (^{15}N or ^{15}N-^{13}C) amino acids during cultivation has been shown to ease the isolation procedures by tracing the metabolites through selective NMR experiments. The latter was termed genomisotopic approach and has been utilized to isolate metabolites like orfamide (Gross et al., 2007). A further possibility is the inactivation of the respective gene cluster and subsequent comparison of the metabolic profile between the wild type and the mutant strains. This should clearly identify the respective compound. In case of silent gene clusters, the genes involved in the biosynthesis of a putative compound may be transformed from the parent strain and expressed in another, possibly optimized host (Wenzel and Müller, 2005). Still another alternative approach is in-vitro reconstitution of the genes and their expression in a heterologous host (Synthetic biology) (Meier and Burkhart, 2009). Also, recombinant biosynthetic proteins can be incubated with their appropriate substrate and the resultant compound is then analyzed further (McClerren et al., 2006). A good option for silent gene clusters is also, to access the regulatory genes thereby activating the transcription of the biosynthetic gene cluster (Bok et al., 2006; Chiang et al., 2011)**.**
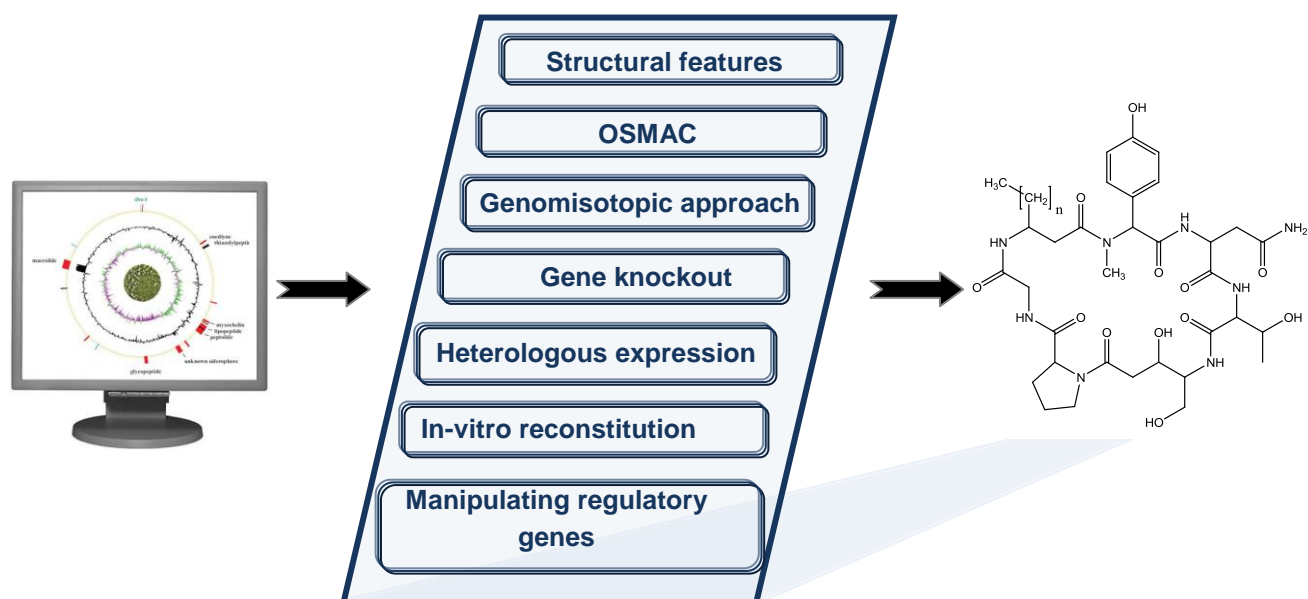
Figure 1-19: Genome mining tools. OSMAC: One strain many compounds. The circular chromosome of *H. aurantiacus* DSM 785 is reproduced from Kiss et al., 2011.

1.12 Genome mining in gliding bacteria

Cyanobacteria have been researched intensely for the isolation of secondary metabolites and are still being considered as a potential source for new secondary metabolites. Genome mining of *Trichodesmium erythraeum* IMS101 revealed a biosynthetic gene cluster encoding for a novel ribosomal peptide. Subsequent cultivation of the bacterium resulted in the isolation of a cyclic peptide, trichamide (Figure 1-20). Indeed this was the first compound known from *T. erythraeum* (Sudek et al., 2006). In addition to this three PKS, two NRPS and a PKS/NRPS hybrid cryptic gene loci were identified from this genome (Kalaitzis et al., 2009). *Nostoc punctiforme* PCC7120 with a genome of 9 Mb in size contributes 3.1% of its genome for the biosynthesis of polyketides and nonribosomal peptides (Kalatzis et al., 2009). Thus, *N. punctiforme* has been identified as the most prolific natural product producer among other sequenced cyanobacteria. In total, 23 PKS and/or NRPS gene loci were identified with a majority of them belonging to PKS/NRPS hybrid systems. A PKS/NRPS hybrid gene cluster has been reported to encode for nostopeptolide biosynthesis (Kalatzis et al., 2009). *Anabaena variabilis* ATCC29413 and *Nostoc* sp. PCC7120 each with a genome of 7 Mb has 2.2% and 1.3% PKS and NRPS genes. All 24 gene loci identified from *An. variabilis* ATCC29413 and *Nostoc* sp. PCC7120 were unknown and thus they are cryptic gene clusters (Kalatzis et al., 2009; Donadio

et al., 2007)). *Synechocystis* sp. PCC6803 which is genetically feasible and easily culturable has already been manipulated for its use as a heterologous host (Roberts et al., 2009). This recombinant strain is useful for the expression of cryptic gene clusters from other cyanobacteria.

Myxobacteria are well known producers of unprecedented structes with biological properties (Wenzel and Müller, 2007; Wenzel and Müller, 2009a). The genome sequences of *Myxococcus xanthus* DK 1622, *Sorangium cellulosum* So ce 56 and *Stigmatella aurantiaca* DW4/3-1 and three *Anaeromyxobacter* species (*Anaeromyxobacter dehalogenans* 2CP-1, *Anaeromyxobacter sp. K, Anaeromyxobacter sp.* FW 109-5) were analysed in order to gain further insight into the potential metabolic profile of the respective bacterial strains (Wenzel and Müller, 2009b). The genomes of *Anaeromyxobacter* species comprise approximately 5 Mb and were reported to be devoid of PKS and NRPS genes, and indeed so far no secondary metabolites are reported from the latter (Wenzel and Müller, 2009b). For *So. cellulosum* Soce56, *M. xanthus* DK1622 and *Sg. aurantiaca* DW4/3-1 biosynthetic genes for specific natural products were already annotated (Gerth et al., 2003; Schneiker et al., 2007; Silakowski et al., 2001; Pradella et al., 2002; Wenzel and Müller, 2009b).

M. xanthus DK1622 which was not known for secondary metabolites (although natural products had been obtained from other *M. xanthus* strains), was the first myxobacterium to have its genome sequenced. A genome size of 9.14 Mb urged to scan it for the presence of biosynthetic genes, since large genomes correlate often with the potential for secondary metabolite production. The sequences revealed almost 8.6% of the genome to possibly be involved in the biosynthesis of secondary metabolites (Krug et al., 2008). Mining the genome for the presence of PKS and NRPS genes exposed 19 biosynthetic clusters, with a predominance of PKS/NRPS hybrid gene clusters. Five of these gene clusters were related to known myxobacterial metabolites, i.e. myxochelin A, myxovirescin, myxochromide, DKxanthenes and myxalamide. Recently, myxoprincomide a novel peptide/polyketide hybrid product was identified by gene inactivation studies (Figure-1.20). Due to the production of myxoprincomide in low amounts, the compound was successfully isolated from another producer strain (Cortina et a., 2011).

So. cellulosum 56 harbors the biggest genome amongst bacteria comprising about 13Mb (Pradella et al., 2002). Since the genus *Sorangium* is noted for the production of secondary metabolites (Gerth et al., 2003), the remarkable genome size of *So. cellulosum* 56 is not surprising (Schneiker et al. 2007). The genome harbours four hybrid PKS/NRPS cluster, one of which being extraordinary huge in size (92.4 kb), and two NRPS as well as three PKS cluster. The products of three of these clusters were attributed to the compounds chivosazole A, myxochelin B and etnangien already isolated from *So. cellulosum*. The putative structures encoded by the rest of the cryptic gene clusters were already predicted and reported (Schneiker et al., 2007). Furthermore, two type III PKS gene loci were reported from the genome. Flaviolin was already isolated by expressing of one of the type III PKS genes in *Pseudomonas putida* (Li and Müller, 2009) (Figure 1-20)

Sg. aurantiaca DW4/3-1 with a genome of 10.3 Mb in size is known for the production of several metabolites namely, myxochelin B, myxochromide S_1-S_3, myxothiazol A, DKxanthenes, aurafuron and dawenol. In addition, the genome sequences revealed four NRPS, four PKS and four PKS/NRPS hybrid cryptic gene clusters. Except dawenol, other compounds were identified and isolated based on the genetic information (Kunze et al., 2005; Meiser et al., 2008; Wenzel and Müller, 2009b) (Figure -1.20)

Chitinophaga pinensis DSM 2588, the first organism to be completely sequenced in the family Chitinophagacae harbors a genome of 9.1 Mb in size. From the genome sequence, a PKS gene cluster was reported to be responsible for the biosynthesis of novel polyketides, elansolid B_1 and D (Teta et al., 2010) (Figure-1.20). Additionally, a cryptic NRPS gene cluster was also reported from the genome sequence (Teta et al., 2010).

Flavobacterium johnsoniae from the phylum Bacteroidetes is noted for its ability to utilize chitin, cellulose and lignin. From the 6.1 Mb genome sequence of *F. johnsoniae* UW-101, an NRPS gene cluster spanning approximately 70.7 Kb was already described. A peptidic structure was predicted fitting the architechture of the 14 NRPS modules but the peptide is not yet isolated (McBride et al., 2009).
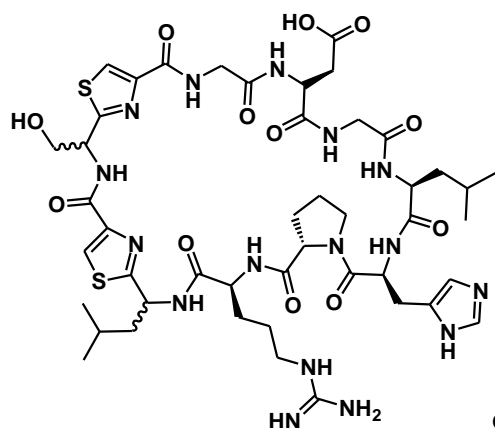
Trichamide- Ribosomal
Structural features
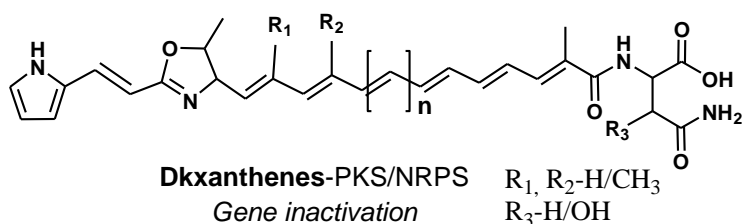
Dkxanthenes-PKS/NRPS
Gene inactivation

R_1, R_2-H/CH_3
R_3-H/OH

Myxovirescin- PKS/NRPS
Gene inactivation
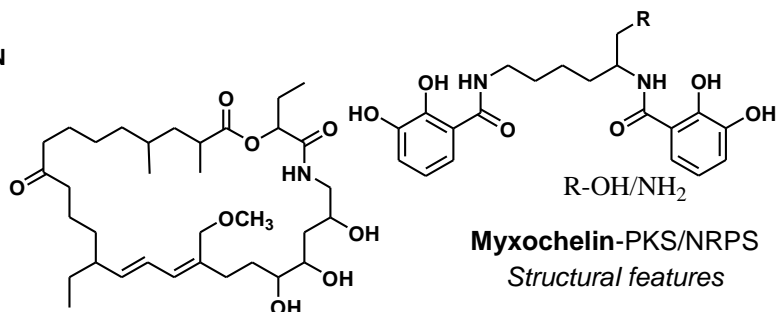
Myxochelin-PKS/NRPS
Structural features

R-OH/NH_2
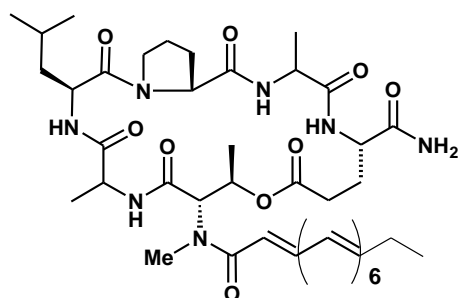
Myxochromide A-PKS/NRPS
Gene inactivation
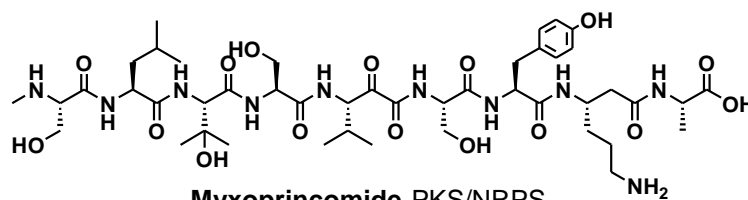
Myxoprincomide-PKS/NRPS
Gene inactivation
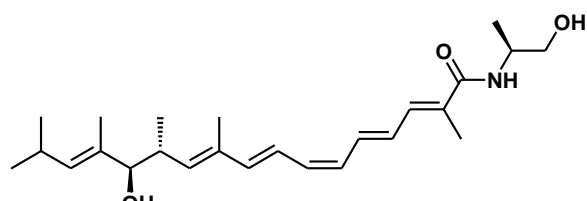
Myxalamide B- PKS/NRPS
Gene inactivation

Flaviolin-PKS type-III
Heterologous expression
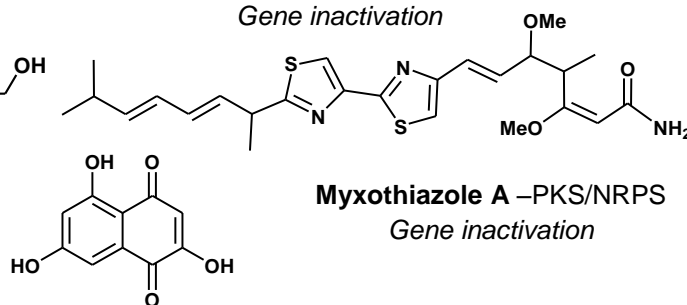
Myxothiazole A –PKS/NRPS
Gene inactivation
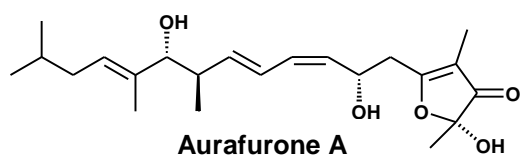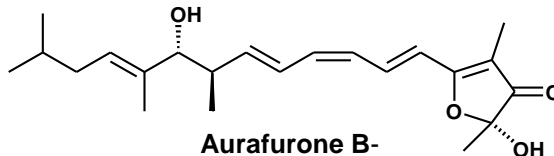
Aurafurone A

Aurafurone B-

PKS-*Gene inactivation*
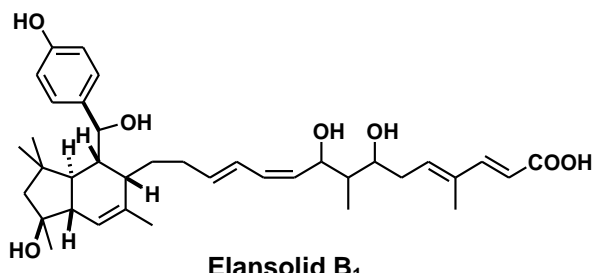
Elansolid B$_1$
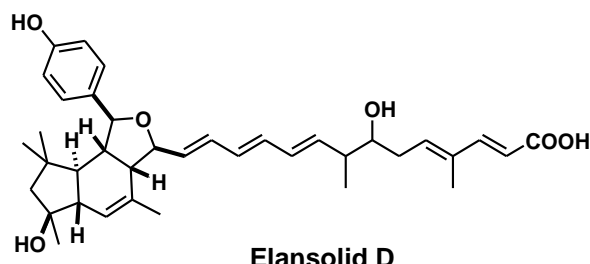
Elansolid D

PKS-*Structural features*

Figure 1-20: Genes to products. Compounds identified from the genomes of gliding bacteria. Mode of identification or isolation from culture extracts are indicated in italics.

Thus, with the advancement in genomics the genetically encoded biosynthetic potential of gliding bacteria becomes even more evident, and bacteria which have not been completely or not at all explored can be considered for chemical investigation based on genomic data. Especially the gene clusters encoding enzymes for the biosynthesis of nonribosomal peptides and polyketides are in most cases readily identified from genome sequences (Donadio et al., 2007).

2. Scope of the project

The advancement in the computational tools and comprehensive knowledge in nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) modular genes has been instrumental in the identification of new biosynthetic gene clusters from genomes. These biosynthetic clusters are often cryptic, i.e. associated with as yet unknown compounds. Thus, this approach called "genome mining" can be applied for the discovery of natural products useful for drug development.

Gliding bacteria, which are known for their complex life style, include the chemically prolific myxobacteria and cyanobacteria. Other bacterial taxa in this group are hardly investigated for their chemistry. Therefore, the first part of this project aimed to evaluate 18 completely sequenced gliding bacterial genomes from the phyla Chloroflexi, Firmicutes, Bacteroidetes, Planctomycetes, Chlorobi and Proteobacteria for the presence of PKS and NRPS genes.

The focus of the second part of the study was a detailed investigation of the complete genome sequence of the chemically unexplored *Herpetosiphon aurantiacus* DSM 785 for PKS and NRPS gene clusters. Information on sequences of cryptic gene clusters and their encoded putative structures were to be used to optimize the conditions for the isolation of secondary metabolites from this bacterium.

In the last part of this study, PKS/NRPS gene locus-1 from *H. aurantiacus* was to be characterized in detail. Emphasis was to be put on the adenylation domains, an integral component of NRPS, which control the entry of amino acid building blocks.

3. Materials and methods

3.1 Materials

3.1.1 Chemicals and solvents

Chemicals	Manufacturer
4-hydroxyphenylglycine	Sigma Aldrich (Germany)
Acetic acid	KMF Laborchemie Handels GmbH (Darmstadt, Germany)
Aceton-d 99.8% D	Deutero GmbH (Kastellaun, Germany)
Acetonitrile	VWR International GmbH (Darmstadt, Germany)
Agar	Fluka Chemie GmbH (Buchs, Switzerland)
Ammonium acetate	Roth Chemie GmbH (Karlsruhe, Germany)
Ammonium per sulfate	Roth Chemie GmbH (Karlsruhe, Germany)
Bacto™ Casitone	Becton, Dickinson and Company (Franklin Lakes, NJ, USA)
Bis-acrylamide	Roth Chemie GmbH (Karlsruhe, Germany)
Boric acid	Roth Chemie GmbH (Karlsruhe, Germany)
Brilliant Blau R 250	Roth Chemie GmbH (Karlsruhe, Germany)
$CaCl_2 \times 2\ H_2O$	Merck KGaA (Darmstadt, Germany)
Chloroform	Roth Chemie GmbH (Karlsruhe, Germany)
DMSO	Roth Chemie GmbH (Karlsruhe, Germany)
dNTPs	Promega GmbH (Mannheim, Germany)
Ethanol 99,8% p.a.	Roth Chemie GmbH (Karlsruhe, Germany)
Ethidium bromide	Roth Chemie GmbH (Karlsruhe, Germany)
Ethyl acetate	Julius Hoesch GmbH & Co. KG (Düren, Germany)
Gel loading dye (6x)	Fermentas GmbH (St. Leon Rot, Germany)
Glycerol	Roth Chemie GmbH (Karlsruhe, Germany)
Imidazole	Roth Chemie GmbH (Karlsruhe, Germany)

Isopropanol	Roth Chemie GmbH (Karlsruhe, Germany)
Isovanillic acid	Alfa Aesar (Ward Hill, MA, USA)
Methanol	Merck KGaA (Darmstadt, Germany)
$MgCl_2 \times 6\ H_2O$	Merck KGaA (Darmstadt, Germany)
$MgSO_4 \times 7\ H_2O$	Merck KGaA (Darmstadt, Germany)
Na_2-EDTA	Roth Chemie GmbH (Karlsruhe, Germany)
Na-acetate	Merck KGaA (Darmstadt, Germany)
NaCl	Merck KGaA (Darmstadt, Germany)
NaOH	Merck KGaA (Darmstadt, Germany)
Ni-NTA agarose	Qiagen GmbH (Hilden, Germany)
peqGOLD Agarose	PEQLAB Biotechnologie GMBH (Erlangen, Germany)
Phenol	Merck KGaA (Darmstadt, Germany)
Polygoprep 60	Macherey-Nagel (Germany)
Roti-Load1 4x Konz	Roth Chemie GmbH (Karlsruhe, Germany)
SDS	Roth Chemie GmbH (Karlsruhe, Germany)
Sephadex LH-20	Pharmacia Biotech (Sweden)
Silica gel 60	Merck KgaA (Darmstadt, Germany)
Sorbitol	Roth Chemie GmbH (Karlsruhe, Germany)
Trifluoracetic acid	Roth Chemie GmbH (Karlsruhe, Germany)
Tris	Roth Chemie GmbH (Karlsruhe, Germany)
Tris-HCl	Roth Chemie GmbH (Karlsruhe, Germany)
Tryptone	Roth Chemie GmbH (Karlsruhe, Germany)
Yeast extract	Fluka Chemie GmbH (Buchs, Switzerland)

Table 3-1: List of chemicals used in this work.

Deuterated solvents used for NMR measurements were obtained from Deutero GmbH (Kastellaum, Germany). Acetone, chloroform, dichloromethane, ethyl acetate and methanol were distilled prior to use. Ultra-pure water for HPLC and molecular

biology experiments was obtained from Milli-Q Academic Water Purification System (Millipore GmbH, Germany). All amino acids used in this study were either from Merck KGaA (Darmstadt, Germany), Sigma Aldrich (Germany) or Roth Chemie GmbH (Karlsruhe, Germany).

3.1.2 Antibiotics

Antibiotic	Manufacturer
Ampicillin	Roth Chemie GmbH (Karlsruhe, Germany)
Gentamycin	Fluka Chemie GmbH (Buchs, Switzerland)
Kanamycin	Sigma-Aldrich Co. LLC (St. Louis, MO, USA)
Streptomycin	Sigma-Aldrich Co. LLC (St. Louis, MO, USA)

Table 3-2: List of antibiotics used in this study

3.1.3 Enzymes

Enzyme	Manufacturer
GoTaq Flexi DNA Polymerase	Promega (Mannheim, Germany)
Pfu DNA Polymerase	Fermentas GmbH (St. Leon-Rot, Germany)
Lysozyme	Roth (Karlsruhe, Germany)
Proteinase K	Roth (Karlsruhe, Germany)
Restriction enzymes	Fermentas GmbH (St. Leon-Rot, Germany)
RNase (DNase free)	Roth Promega (Mannheim, Germany)
DNase	Fermentas GmbH (St. Leon-Rot, Germany)
T4 DNA-ligase	Fermentas GmbH (St. Leon-Rot, Germany)

Table 3-3: List of enzymes used in this study

3.1.4 Kits and Standards

Article	Manufacturer
DNA Clean & ConcentratorTM	Zymo Research Corporation (Irvine, USA)
Gene RulerTM 1kb plus DNA ladder	Fermentas GmbH (St. Leon-Rot, Germany)
Gene RulerTM 50bp DNA ladder	Fermentas GmbH (St. Leon-Rot, Germany)
Gene RulerTM DNA Ladder Mix	Fermentas GmbH (St. Leon-Rot, Germany)
Gene RulerTM high range DNA ladder	Fermentas GmbH (St. Leon-Rot, Germany)
PureYieldTM Plasmid Miniprep System	Promega (Mannheim, Germany)
QIAGEN Plasmid Maxi Kit	Qiagen GmbH (Hilden, Germany)
QIAGEN Plasmid Midi Kit	Qiagen GmbH (Hilden, Germany)
QIAprep Spin Miniprep Kit	Qiagen GmbH (Hilden, Germany)
QIAquick Gel Extraction Kit	Qiagen GmbH (Hilden, Germany)
QIAquick PCR Purification Kit	Qiagen GmbH (Hilden, Germany)
RevertAid H Minus First Strand cDNA Synthesis Kit	Fermentas GmbH (St. Leon-Rot, Germany)
Unstained Protein Standard	Fermentas GmbH (St. Leon-Rot, Germany)
Wizard SV Gel and PCR Clean-Up System	Promega (Mannheim, Germany)

Table 3-4: List of kits and standards used in this study

3.1.5 Solutions for DNA Analysis

Name	Constituents and their amounts
Buffer P1	50 mM Tris-HCl (pH 8), 10 mM EDTA, 100 µg/mL RNase A
Buffer P2	200 mM NaOH, 1% SDS
Buffer P3	3M potassium acetate (pH 5.5)
Proteinase K solution	1 g Proteinase K was added to 50 ml DNAase/RNAse free water, aliquoted and stored at -80°C.

RNase Solution	500 mg RNase was dissolved in 25 ml water and heated for 15 min at 95°C. Aliquoted and stored at -80°C.
TBE (10 x)	0.89 M Tris base, 0.02 M EDTA, 0.87 M boric acid, pH 8.3.
TE buffer	10mM Tris-HCl (pH 8.0), 1mM EDTA

Table 3-5: List of buffers and solutions prepared for DNA analysis

3.1.6 Solutions for Protein Biochemistry

Name	Constituents and their amounts
Lysis buffer	50 mM NaH2PO4, 300 mM NaCl, 10 mM imidazole, pH 8.0
Wash buffer	50 mM NaH2PO4, 300 mM NaCl, 20 mM imidazole, pH 8.0; for more stringent washing 40 mM, 60 mM and 100 mM imidazole were applied
Elution buffer	50 mM NaH2PO4, 300 mM NaCl, 300 mM imidazole, pH 8.0
SDS-running buffer (10x Laemmli buffer)	250 mM Tris, 2M glycine, 1% SDS, pH 8.9
Coomassie staining solution	1.0 g Coomassie Brilliant Blue R-250, 25 ml acetic acid, 200 ml methanol and the volume was adjusted to 500 ml with distilled water and stored at room temperature. Solution was used more than once.
Coomassie destaining solution	50 ml acetic acid, 150 ml methanol and the solution made up to 500 ml with water.
A Domain buffer	20 mM Tris-HCl (pH 7.5), 5% glycerol

Table 3-6: List of buffers and solutions prepared for protein analysis

3.1.7 Media

Medium	Composition
LB medium	10 g tryptone, 5 g yeast extract, 10 g NaCl in 1 L of water, pH 7.5
LB agar	10 g tryptone, 5 g yeast extract, 5 g NaCl, 15 g Agar, in 1 L of water, pH 7.5
2 x YT medium	16 g tryptone, 10 g yeast extract, 5g NaCl in 1 L of water, pH-7.5

Trace element Solution	20 mg $ZnCl_2$, 100 mg $MnCl_2$ x $4H_2O$, 10 mg boric acid, 10 mg $CuSO_4$, 20 mg $CoCl_2$, 5 mg $SnCl_2$ x $2H_2O$, 5 mg LiCl, 20 mg KBr, 20 mg KI, 10 mg Na_2MoO_4 x $2H_2O$ and 5.2 g Na_2-EDTA x $2H_2O$ in 1 L of water. The solution was sterilized by filtration.
CY medium	3.0 g casiton, 1.36 g CaCl2 x 2 H2O; 1.0 g yeast extract in 1 L of water, pH 7.2; after autoclaving add 1 ml filter sterilised trace element solution and 1 ml vitamin B12 solution (0.5 mg/ml)
VY/2 agar	50 ml sterilised baker's yeast suspension (10%), 1.36 g CaCl2 x 2 H2O, 15 g agar in 1 L of water, pH 7.2; after autoclaving add 1 ml filter sterilised vitamin B12 solution (0.5 mg/ml)
MD1+G	3 g casitone, 0.7g $CaCl_2$ x $2H_2O$, 2 g $MgSO_4$ x $7H_2O$ in I L of water. The autoclaved medium was then supplemented with 0.2 % glucose and sterile filtered cyanocobalamine (0.5 mg/L) and trace element solution (1 ml/L)
HP74	10 g sodium glutamate, 2 g yeast extract, 2 g $MgSO_4$ x $7H_2O$ in 1 of water. The autoclaved medium was then supplemented with 1% glucose.

Table 3-7: List of media used for cultivating bacteria in this study

3.1.8 Organisms

Organism	Genotype	Provider
Herpetosiphon aurantiacus DSM 785	wild type	DSMZ
XL1-Blue *E. coli* cells	recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F′ proAB lacIqZΔM15 Tn10 (Tetr)]	Stratagene (La Jolla, CA, USA)
BL21 (DE3)pLysS	F– ompT gal dcm lon hsdSB(rB- mB-) λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])	Invitrogen Life Technologies Corporation (Karlsruhe, Germany)
Rosette (DE3)	F$^-$ *ompT hsdS*$_B$(r$_B^-$ m$_B^-$) *gal dcm* λ(DE3) pRARE (Cam^R)	Invitrogen Life Technologies Corporation (Karlsruhe, Germany)
ArcticExpress (DE3)	*E. coli* B F– *ompT hsdS* (rB– mB–) *dcm*+ Tetr *gal* λ(DE3) *endA* Hte [*cpn10 cpn60* Gentr] [*argU proL* Strr]	Agilent Technologies (Santa Clara, CA, USA)

Table 3-8: List of bacteria used in this study

3.1.9 Vectors

Vector	Resistance	Manufacturer
pET28a(+)	kanamycin	Merck KGaA (Darmstadt, Germany)
pGEM-T	ampicillin	Promega (Mannheim, Germany)
pCDF-Duet	Streptomycin	Dr. Max Crüsemann, Institute for Organic Chemistry, Bonn

Table 3-9: List of vectors used in this study

3.1.10 Softwares and databases

Genome sequences used for analysis in this study were accessed from GenBank (http://www.ncbi.nlm.nih.gov). These sequences were then scanned for the presence of PKS and NRPS genes (Jenke-Kodama and Dittmann, 2009) using Cluster Sequence Analyzer (CLUSEAN). A graphical representation of PKS and NRPS domain organization was obtained from amino acid sequences by freely available databases such as PKS/NRPS analysis (Bachmann and Ravel, 2009) and NRPS/PKS database (Ansari et al., 2004). The specificity of the A domains in an NRPS system was identified by a specialized tool called NRPS predictor (Rausch et a., 2005) and NRPSpredictor2 (Rottig et al., 2011). The specificity of AT domains were analysed based on Yadav et al., (2003) and Smith and Tsai (2007) by mulitple sequence alignment using CLUSTALW (Goujon et al., 2010; Larkin et al., 2007) with similar AT domains obtained from BLAST. All domains including the accessory domains were analyzed for their activity by identifiying the catalytically active residues in their amino acid sequences by BLAST and CLUSTALW. The nucleotide sequences used for cloning experiments were based on the amino acid regions specified by Pfam.

Program resource	Hyperlink	Enzymes	Ref
CLUSEAN	http://redmine.secondarymetabolites.org/projects/clusean	NRPS, PKS	Weber et al., 2009
PKS/NRPS analysis	http://nrps.igs.umaryland.edu/nrps/	NRPS, PKS	Bachmann and Ravel, 2009
NRPS/PKS database	http://www.nii.res.in/searchall.html	NRPS, PKS	Ansari et al., 2004
NRPSpredictor	http://ab.inf.uni-	NRPS	Rausch et al., 2005;

	tuebingen.de/toolbox/index.php?view=domainpred		Rottig et al., 2011.
NRPSPredictor2	http://nrps.informatik.uni-tuebingen.de/Controller?cmd=SubmitJob		

Table 3-10: List of tools used in this study to analyse the genome sequences for PKS and NRPS genes.

3.1.11 Primers

All primers used in this study were from Eurofins MWG Operon (Ebersberg, Germany) as salt free lyophilized water. Once received, the oligonucleotides were suspended in sterile water to a concentration of 100 pmol/µl and stored at -20°C.

Gene locus	Primer	Sequence (5'-3')
Primers for gene expression analysis		
haur_1857	P_1	GATCTGAGCAGCCAACAC CTGTCAGCACGGCTTTAG
haur_1858-59	P_2	TGTTATAGCACTGATGGCTG CTTTGTAATTGGCTGGTTTC
haur_1860-61	P_3	CTTAATTATTCAGGCCGATG AGCCACAATTTTATGTCGAG
haur_1871	P_4	CGCCGTGTTGCACTGCCAAT AGTTGAAGCTACGCCCGCGA
haur_1887-88	P_5	GCAAATCGCGCAGCATCGCA AAACGTGGTGCCCGCCCAAT
haur_2089	P_6	ACGACCAGCCATCAAAAATC ATGTCGCATACTTGCAGCAG
haur_2091	P_7	CCCGCTCCAACAATTGAC ATGGCTGGTCGATTGGAA
haur_2108	P_8	AGGCTTGCTCACCAAACG GATCAGGGCATTGCCAAC
haur_0810	P_9	CTATAACTCGGGCTTGATTG CACAATCACAATTTGCTCAC
Primers for protein expression analysis		
haur_1858-A_3	P_{10}	<u>ATTTAT</u>**GGATCC**GGGCTAGTTGGGCGCTTG <u>TATAT</u>**CTCGAG**TTAGTGGGGCAACTGTTGGGC
haur_1859-A_4	P_{11}	<u>GATCGT</u>**CATATG**AGCTACGCCGAGCTTGAT <u>ATATTA</u>**GAATTC**CACCGCCGCCGATTGCAC
haur_1859-A_5	P_{12}	<u>GATCGA</u>**CATATG**AGCTATGCCGAGCTGGAT <u>GGCCGC</u>**GAATTC**TACAACGATTTGCGCAAT
haur_1857-A_2-PCP_2	P_{13}	<u>CTCG</u>**GGATCC**ATGCGAAGATTACTCCTTGGC GCGGCCGTATACAACATCTCTTTGATCG

haur_1858-A$_3$-PCP$_3$	P$_{14}$	<u>ATTTAT</u>**GGATCC**GGGCTAGTTGGGCGCTTG <u>TGTGTCTCGAGTTAAATCGCTTGGGCCAAGCT</u>
haur_3130-PPT	P$_{15}$	<u>CGTGGTTC</u>**CATATG**ATCCTATGTTCTATG <u>AGCTAGC</u>**AAGCTT**ACTCAGCCAAGCCCAG
haur_2109-mbtH	P$_{16}$	<u>GCGAGT</u>**CATATG**GTGCAAACAACAGATGAC <u>AATTAT</u>**CTCGAG**CTAAGCCGATTGGTTGGC

Table 3-11: List of primers used in this study. Junk sequences for the purpose of cloning are underlined and restriction sites are highlighted in bold.

3.1.12 Constructed recombinant plasmids

Construct	Plasmid	Description
pLN-A$_3$	pET28a(+)	*Haur_1858-A$_3$* amplified from *H. aurantiacus* genomic DNA by P$_{10}$ primers.
pLNA$_4$	pET28a(+)	*Haur_1859-A$_4$* amplified from *H. aurantiacus* genomic DNA by P$_{11}$ primers.
pLNA$_5$	pET28a(+)	*Haur_1859-A$_5$* amplified from *H. aurantiacus* genomic DNA by P$_{12}$ primers.
pLNA$_3$-PCP$_3$	pET28a(+)	*Haur_1858-A$_3$-PCP$_3$* amplified from *H. aurantiacus* genomic DNA by P$_{14}$ primers.
pLNPPT	pET28a(+)	*Haur_3130* amplified from *H. aurantiacus* genomic DNA by P$_{15}$ primers.
pLNA$_2$-PCP$_2$	pET28a(+)	*Haur_1857-A$_2$-PCP$_2$* amplified from *H. aurantiacus* genomic DNA by P$_{13}$ primers.
pLN-*mbtH*	pCDF-Duet	*Haur_2109* amplified from *H. aurantiacus* genomic DNA by P$_{16}$ primers.
RP$_{1-9}$	pGEM-T	*H. aurantiacus* cDNA amplified with respective primer pairs P$_1$-P$_9$.

Table 3-12: List of recombinant plasmids constructed in this study

3.2 Microbiological methods

3.2.1 Sterilization

All microbiology and molecular biology experiments were carried out in a sterile environment. To accomplish this, materials and solutions were steam sterilized before use by autoclaving at 121°C for 20 min in Varioklav steam sterilizer. In case of

heat sensitive solutions such as antibiotics, IPTG and culture media supplements sterilization was performed by filtration using 0.22 µm membrane filters.

3.2.2 Preserving stock cultures

For long term storage of the *Escherichia coli* strains, 900 µl of overnight culture was mixed with 600 µl of 80% glycerol in a cryovial and stored at -80°C. *Herpetosiphon aurantiacus* DSM 785 stocks were prepared by adding 500 µl of 7 days old liquid culture with equal volume of 2% casitone in cryovials and stored at -80°C.

3.2.3 Cultivation of *E. coli*

E. coli cultures were grown aerobically in LB medium in the presence of required antibiotic (Table 3-2). Unless described, the cultures were incubated overnight at 37°C and 180 rpm. The cell density was monitored by measuring the optical density at 600 nm (OD_{600}) in UV mini 1240 UV/Vis spectro-photometer.

3.2.4 Cultivation of *H. aurantiacus*

H. aurantiacus was cultivated in order to screen the bacterium for the presence of secondary metabolites. To explore the metabolic profile of *H. aurantiacus* different media were used for cultivation. They include Casiton-Yeast (CY) based medium, MD1+G and HP74 medium. However, in all conditions agar block of .*H. aurantiacus* grown in VY/2 agar at 30°C for 7 days was used to cultivate 50 ml liquid pre-culture. The preculture, which was grown at 30°C and 140 rpm for 7 days, was used as inoculum for 1.5 L culture. The main culture was also allowed to grow for 7 days before extraction.

To the main culture in CY medium, supplements were added at different time points to induce the production of secondary metabolites. In one set of culture in CY medium, 2 mM of filter sterilized tyrosine was added after 2 days. In other set of cultures in same medium, filter sterilized 3 mM 4-hydroxyphenylglycine was added. *H. aurantiacus* lyses dead *E. coli* cells for nutrition and therefore, 10 ml of dead *E. coli* cells was added to 1 L of 3 days old culture in same medium and incubated further for 4 days. The dead *E. coli* cells were made by autoclaving 50 ml suspension obtained from the lawn of live *E. coli* cells in LB agar plate.

3.3 Chemical methods

3.3.1 Crude extraction of bacteria

3.3.1.1 Cell pellet

To the cell pellet, 100 ml methanol was added and stirred overnight at room temperature. The methanol extract was then centrifuged at 8000 rpm for 15 min at room temperature. The supernatant was evaporated under vacuum for further fractionation and screening.

In other case, the cell pellet was lyophilized in freeze dryer. Lyophilized biomass was suspended in 200 ml of methanol and dichloromethane (2:1) mixture and stirred at room temperature for 30 min. The suspension was then centrifuged at 8000 rpm for 15 min. The supernatant was collected and the cell pellet was extracted again exhaustively. The collected supernatant were pooled together and evaporated under vacuum for further procedures.

3.3.1.2 Supernatant

In the first method, the supernatant was extracted exhaustively with equal volume of either ethyl acetate or butanol by liquid–liquid partition. The organic phase was collected and evaporated under vacuum. The extract was then further fractionated and all the fractions were subjected to LC-MS for screening.

The second method, which involves alternate acidification and neutralization, has been successful in isolating lipopeptides like laspartomycin and aspartocin (Borders et al., 2004). According to this procedure butanol phase was initially mixed with equal volume of H_2O and adjusted to a pH-2 using 1N HCl. Then the acidified butanol phase was washed with water and mixed with equal volume of water. This mixture was neutralized to pH-7 and the aqueous phase was separated. This aqueous phase was again acidified to pH-2 and extracted with butanol. The aqueous phase was adjusted back to pH-7 and should contain lipopeptides free of impurities.

Finally, the extraction procedure used for the isolation of lipopeptide surfactin (Ohno et al.,1995) was followed. As per this procedure, the supernatant was adjusted to pH-2 using 2 N hydrochloric acid. The mixture was incubated overnight at 4°C to allow

any lipopeptides to precipitate. The precipitate was then collected by centrifuging at 5000 rpm at 4°C for 45 min. The precipitate was extracted with 100 ml of methanol for 1 h. The methanol extract was centrifuged at 8,000 rpm at 4°C for 15 min and the supernatant was evaporated under vacuum.

3.3.2 Chromatography

The obtained crude extracts were further fractionated by different chromatographic techniques to isolate the pure compound. They include gel filtration chromatography, vacuum liquid chromatography (VLC), and high performance liquid chromatography (HPLC)

3.3.2.1 Gel filtration and vacuum liquid chromatography (VLC)

Gel filtration chromatography was carried out in Sephadex LH-20 (0.018-0.111 mm, Pharmacia Biotech AB; size exclusion material), using MeOH as eluent. The column was wet packed with MeOH before applying the sample solution. Sample was also loaded by using wet method. 20 ml of each fraction was collected, evaporated at 40°C in a rotary evapaorator.

VLC was carried out using Silical gel 60 (0.040-0.063 mm, Merck 1.09385.1000) or silica gel 60 (0.063-0.200 mm, Merck 1.07734.2500) as the stationary phase for normal phase (NP), whereas polygoprep 60 C_{18} (0.05 mm, Macherey-Nagel 711500.1000) for reverse phase-vacuum liquid chromatography (RP-VLC). The columns were wet-packed with stationary phase soaked in dichloromethane for NP-VLC or in methanol for RP-VLC. Sample was loaded by wet method Dichloromethane, ethyl acetate, acetone and methanol were used as mobile phase in NP-VLC, whereas in RP-VLC gradient of water and methanol were used as mobile phase. Fractions were collected and evaporated at 40°C under reduced pressure on a rotary evaporator.

Prepacked Bakerbond Spe C_{18} disposable extraction columns (J. T. Baker Inc. Phillipsburg, NJ, USA; Cas No. 71889-02-06) were used for the fractionation of samples.

3.3.2.2 High performance liquid chromatography (HPLC)

HPLC was carried out in Waters system. The system was equipped with in-line degasser, 717 plus autosampler, 600 pump and 996 photodiode array detector. Columns used were either Knauer Eurospher-100 C_{18} (5 μm, 250 x 8 mm), Knauer Eurospher-100 C_8 (5 μm, 250 x 8 mm) or similar columns. Milli-Q purified water, distilled methanol and HPLC grade acetonitrile were used as mobile phase. Typical flow rates were 1.0-2.5 mL/min.

3.4 Analytical methods

3.4.1 NMR spectroscopy

NMR analyses of the extracts were either carried out on a Bruker Avance 300DPX spectrometer operating at 300 MHz for ^{1}H and 75 MHz for ^{13}C or Bruker Avance 500 DPX spectrometer operating at 500 MHz for ^{1}H and 125 MHz for ^{13}C. The spectra were processed using XWIN NMR Version 3.1 software. Deuderated solvents were used to suspend the fractions for NMR measurement.

3.4.2 Mass spectrometry

LC-MS measurements were conducted by E. Eguereva, Institute for Pharmaceutical Biology, Bonn, Germany. The samples were dissolved in methanol (1 mg/mL) for injection into the system. Experiments were recorded in both positive (+Q1) and negative mode (-Q1) and analysed by Applied Biosystems/ MDS Sciex Analyst software.

HR-MS of the crude extract was performed by Dr. M. Nett in Leibniz Institute for Natural Product Research and Infection Biology, Jena, Germany. The sample was dissolved in methanol (1.5 mg/ml). Experiment was recorded in both positive (+Q1) and negative (-Q1) modes.

In-gel trypsin digestion of protein bands of interest from a native-PAGE and subsequent analysis of the mass and peptide fingerprint was carried out in Bruker MALDI TOF/TOF mass spectrometer by Dr. Sebastian Franken and Dr. Till Schäberle in Institute for Biochemistry and Molecular Biology, Bonn, Germany.

γ-$^{18}O_4$-ATP pyrophosphate exchange assays was measured using MALDI-TOF MS by Dr. Max Crüsemann, Institute for Organic Chemistry, Bonn. Mass spectra were acquired in negative ion mode over a range of 450 to 1200 m/z. Data acquisition and quantitative spectral analysis was conducted using Bruker Compass Data Analysis software. 1 µl of the assay mixture was used for the measurement.

3.4.3 UV measurements

UV spectra were recorded on a Perkin-Elmer Lambda 40 with UV WinLab Version 2.80.03 software, using 1.0 cm quartz cells.

3.5 Molecular biological methods

3.5.1 Isolation of nucleic acids

3.5.1.1 Chromosomal DNA

Chromosomal DNA was isolated according to Neumann et al.,1992. Bacterial cells were harvested from 100 ml of culture by centrifuging at 8000 rpm for 10 min. All centrifugation steps were performed at 4°C. The cell pellet was resuspended in 5 ml SET buffer. To the suspension 0.5 ml SDS (10%), 275 µl lysozyme (100mg/ml) and 275 µl proteinase K (20 mg/ml) was added and incubated for 2 hours at 55°C with intermittent mixing for every 15 minutes. 1.8 ml 5 M NaCl and an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added to the lysate and gently mixed by inverting several times. The aqueous phase was then separated from the organic phase by centrifuging at 8000 rpm for 10 minutes. To get rid of phenol residues in the aqueous phase, the same step was carried out with chloroform: isoamyl alcohol (24:1).

The aqueous phase containing genomic DNA was mixed with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2-2.5 volumes of ice-cold ethanol 98%. The mixture was incubated at 4°C for 1 h. The precipitated DNA was centrifuged at 8000 rpm at 4°C for 30 min. The supernatant was decanted and the DNA pellet was dried at 37°C. The DNA was redissolved in TE buffer (pH 8.0) containing RNase (20 µg/ml) and was incubated at 37°C for 30 min to degrade the RNA. The RNA free DNA was then stored at 4°C.

3.5.1.2 Plasmid DNA

Plasmid DNA was routinely isolated by alkaline lysis methods (Birnboim and Dolly, 1979). For the small scale experiments, 3 ml of overnight bacterial culture was used to isolate the plasmid DNA. The cells were harvested from the overnight culture by centrifuging at 13,000 rpm for 2 min. The supernatant was decanted and the cell pellet was resuspended in 350 µl of P1 buffer. To the suspension, 300 µl of P2 buffer was added and tube was mixed by inverting six times. Within 2 min, 300 µl of P3 buffer was added and mixed by inverting six times. The suspension was then centrifuged at 13,000 rpm for 10 min. Approximately, 900 µl of the supernatant was then carefully transferred to a new eppendorf without disturbing the pellet. To the supernatant, equal volume of ice-cold 96% ethanol was added. This mixture was then centrifuged at 13,000 rpm for 30 min at 4°C. The supernatant was carefully discarded and the DNA pellet was washed in 70% ethanol. The mixture was again centrifuged for 30 min at 4°C. The supernatant was discarded and the DNA pellet was dried. The pellet was then resuspended in 30-50 µl of TE buffer or H_2O and then stored at -20°C.

For sequencing purposes, highly pure plasmid DNA was isolated using PlamidMini Kit from Promega (Mannheim). For large scale purification of plasmid DNA, QIAGEN Plasmid Midi kit was used as recommended by the manufacturer.

3.5.1.3 RNA and DNase reaction

Extreme care was taken during the RNA isolation to prevent its degradation from the ribonucleases. The work space and all the labwares were cleaned with 0.5 N NaOH before and during isolation to get rid of the ribonucleases. Sterile tips and eppendorf tubes were used while handling RNA.

Total RNA was isolated using Macherey Nagel Nucleospin RNA II kit as per the protocol from the manufacturer. RNA was then checked for any traces of DNA which would otherwise lead to false positives in RT-PCR. DNA free RNA was ensured by doing a PCR (section-3.5.3.1 with 3µl of the template) with the gene specific primers designed for RT-PCR (section-3.5.3.2). Amplified PCR fragments indicated the

presence of DNA. The contaminating DNA in the identified RNA preps was then removed with DNase as follows.

Reagent	Volume (µl)
RNA	40
DNase reaction buffer	90
DNase	10

Table 3-13: DNase reaction mix

The solution was then incubated at 37°C for 1 hr followed by inactivation at 65°C for 10 min.

3.5.2 Evaluation of nucleic acids

3.5.2.1 Agarose gel electrophoresis

Agarose gel electrophoresis is a common method used to separate DNA based on its molecular weight. Standard gels containing 1% electrophoresis grade agarose was used to separate the DNA fragments of size ranging from 200bp to 10Kbp. The electrophoresis was carried out with the running buffer (0.1X TBE) covering the gel. The DNA samples were mixed with 0.2 volumes of 6X loading dye, which makes easy to monitor the progress of the gel. 5-10 µl of 1Kbp DNA ladder was used as standard size indicator. Electrophoresis was done in Horizon-58 and Horizon-11.14 gel chamber (Life Technologies, Karlsruhe, Germany) at a voltage ranged between 80 to 100V. To visualize the DNA, the gel was stained with ethidium bromide solution (10µg/ml) for 2 minutes. Subsequently, non-intercalated ethidium bromide was washed out for 2 minutes in demineralized water. DNA bands were visible under UV (254nm) and documented using intasiX imager (Intas Science Imaging Instruments GmbH, Göttingen, Germany).

3.5.2.2 DNA extraction from agarose gels

DNA fragments separated in the agarose gel was extracted and purified using QIAquick gel extraction kit (Qiagen, Hilden, Germany) as per the manufacturer's instructions. The purified DNA was eluted with 30-50 µl water and stored at -20°C.

3.5.2.3 Determination of nucleic acid concentration and purity

The nucleic acids concentration was determined using UV mini 1240 UV/Vis spectrophotometer (Shimadzu, Kyoto, Japan). In principle as per the Lambert-Beer law, the amount of energy absorbed at a particular wavelength is a function of the concentration of the absorbing material. Usually, the nucleic acids have the maximum absorbance at 260 nm. Apart from the nucleic acids, the impurities from the nucleic acid preps usually absorbs in the range of 200 nm to 320 nm. Therefore, the absorbance at 260 nm can be adjusted for background correction using the absorbance at 320 nm. Using these absorbance values and the dilution factor, the nucleic acid concentration was calculated using the following equation.

$$\text{Nucleic acid concentration (µg)} = (A_{260} - A_{320}) \times \text{extinction coefficient} \times \text{dilution factor}$$

$$A_{260} \longrightarrow \text{Absorbance at 260 nm}$$
$$A_{320} \longrightarrow \text{Absorbance at 320 nm}$$
$$\text{Extinction coefficient } (A_{260} \text{ of } 1.0) \longrightarrow 50 \text{ µg/ml for DNA and 40 µg/ml for RNA}$$

Proteins in particular the aromatic amino acids absorb at 280nm. So, therefore the ratio of the absorbance at 260 nm and 280 nm can be used to determine the purity of the sample. $A_{260} : A_{280}$ in the range of 1.7 to 2.0 represents highly pure nucleic acids.

3.5.3 Amplification of nucleic acids

3.5.3.1 Polymerase Chain Reaction (PCR)

PCR is an in vitro technique used for the amplification of the gene of interest. In general, oligonucleotides (18-20 bp) termed as primer complementary to the regions flanking the gene of interest was designed. These primers then hybridize to the single stranded DNA and then subsequently elongated at the 3' end of the primer.

Reagent	Volume
5X *Taq*-buffer (green)	4.0µl
dNTP mix (10mM)	0.4µl
Forward primer (100pmol)	0.5µl
Reverse primer (100pmol)	0.5µl
DMSO	1.0µl

Template	1.0µl
Taq DNA Polymerase(5 units/µl)	0.2µl
H_2O	12.4µl

Table 3-14: Standard PCR master mix with *Taq* DNA Polymerase

Reagent	Volume
10X *Pfu* buffer with $MgSO_4$	5.0µl
dNTP mix (10mM)	1.0µl
Forward primer (100pmol)	0.5µl
Reverse primer (100pmol)	0.5µl
DMSO	1.0µl
Template	1.0µl
Pfu DNA Polymerase(2.5 units/µl)	1.0µl
H_2O	40.0µl

Table 3-15: Standard PCR master mix with *Pfu* DNA Polymerase

No	Step	Temperature	Time	No. of cycles
1	Initial denaturation	95°C	3 min	1
2	Denaturation	95°C	45 sec	
3	Annealing	$T_m - 5°C$	45 sec	30 cycles
4	Extension	72°C	X sec	
5	Final extension	72°C	15 min	1

Table 3-16: Standard cycling protocol

T_m is the melting point of the primers. The extension time depends on the polymerase enzyme and the length of the fragment to be amplified. *Taq* polymerase lacking in the proof reading ability is used for normal PCR amplifications. For *Taq* polymerase the extension time is 1 min/Kb. *Pfu* polymerase with its proofreading property is usually used for cloning purposes. The extension time is 2 min/Kb.

3.5.3.2 Reverse Transcriptional-Polymerase Chain Reaction (RT-PCR)

The expression of the specific genes in the bacterial genome was analysed using RT-PCR. The isolated RNA was used as a template for the cDNA synthesis using the forward primer of the designed gene specific primers. Reverse transcription of the

RNA was done using the Revertaid^TM first strand cDNA synthesis kit. The reaction was carried out with revertaid reverse transcriptase, reverse transcriptase reaction buffer, riboblock RNase inhibitor, dNTP, RNA and the gene specific primers Table 3-11) as per the instructions from the manufacturer.

With 3µl of the synthesized cDNA as the template, PCR (section-3.5.3.1) was done using the respective gene specific primers. An amplified PCR fragment of the expected size was an indicator for the successful reverse transcription reaction. The RT-PCR reaction was then confirmed by sequencing.

3.5.4 Enzymatic manipulation of DNA

3.5.4.1 Restriction digestion of the DNA

The PCR amplified DNA and the plasmid DNA was digested with the respective restriction endonucleases. The respective site for the restriction enzyme was defined by the recognition sequence in the DNA.

All the restriction enzymes used in this research were from Fermentas. The double digestions were performed as per the instructions from the manufacturer.

Reagent	Plasmid DNA (170ng/µl)	PCR product (60ng/µl)
DNA	3 µl	10 µl
10X Tango buffer	4 µl	6 µl
Restriction enzyme I &II	X µl	X µl
Nuclease-free water	To 20 µl	To 30µl

Table 3-17: Standard restriction digestion reaction mix

The volume of the enzyme varies with the combination of the enzymes used for digestion. No matter, the total volume of the restriction enzymes should be restricted to 10% of the total volume. This is important because the glycerol in the enzyme could interfere with the digestion reaction. The optimal temperature for all the restriction enzymes used was 37°C.The reactions were incubated at 37°C for 16 hrs. The specific fragment was then selected and purified form the agarose gel.

3.5.4.2 Ligation

The restriction digested PCR product and the plasmid DNA was ligated using T4 DNA ligase. The reaction was performed as per the manufacturer's instruction.

Reagent	Volume
Plasmid DNA (100ng/µl)	3 µl
PCR product (50ng/µl)	7 µl
10X T4 DNA ligase buffer	2 µl
T4 DNA ligase (1U/µl)	1 µl
Nuclease-free water	7 µl

Table 3-18: Ligation reaction mix

The reaction was incubated at 16°C for 16 hrs. The reaction was stopped by heating at 70°C for 5min.

In addition, the PCR products generated from the synthesized cDNA was ligated into the pGEM-T vector for sequencing. The amplicons generated with *Taq* DNA polymerase have 3' A- overhangs. So that, the A-tailed amplicons could be directly ligated with the T-Tailed pGEM-T vector by T4 DNA ligase without restriction digestion. The ligation was carried out as per the manufacturer's instruction described earlier. Typically, 1:3 vector to insert ratio was maintained in the ligation reaction.

3.5.5 Bacterial Transformation

3.5.5.1 Preparation of chemically competent cells

E. coli XL 1 blue competent cells were prepared according to the method described by Dagert and Ehrlich, 1974. A single colony from a fresh agar plate was inoculated into 3 ml LB medium. The inoculum was then incubated at 37°C overnight with shaking at 160 rpm. The overnight culture was then inoculated in fresh 70ml 2 X YT medium. The culture was cultivated at 37°C at 160 rpm until it reaches an OD_{600} of 0.3-0.4. Bacterial cells were harvested by centrifuging at 5000 rpm for 10 min. All the centrifugation steps were carried out at 4°C. The bacterial cell pellet was then resuspended in 10 ml of ice-cold 70 mM $CaCl_2$/20 mM $MgSO_4$ solution. The

suspension was incubated in ice for 30 min and centrifuged at 5000 rpm for 10 min. The cell pellet was again resuspended in 3.5 ml ice-cold 70mM $CaCl_2$/20 mM $MgSO_4$ and incubated in ice for another 30 min. To this solution 875µl sterile glycerol was mixed and aliquoted as 100 µl volumes in 1.5 ml eppendorf tubes. For storage, these competent cells were shock frosted in liquid nitrogen and were stored at -20°C.

3.5.5.2 Transforming chemically competent cells

The frozen competent cells were thawed in ice for 5 min. To that 5µl of the plasmid DNA was added and incubated in ice for 30 min. Then the cells were heat shocked at 42°C for 50 sec. The cells were then incubated again in ice for atleast 2 min. To the cold shocked cells 900µl of LB medium was added and incubated at 37°C with shaking at 160 rpm for 1 hr. The cells were then centrifuged at 13,000 rpm for 5 min. The cell pellet was then resuspended in 300 µl of fresh LB medium. 100 µl of the cell suspension was then plated on LB agar plate containing ampicillin (100 µg/ml) as the selection marker. The plate was then incubated overnight at 37°C. The positive transformants was then identified by restriction digestion (section-3.5.4.1), PCR (section-3.5.3.1) and sequencing.

3.5.6 Recombinant protein expression

3.5.6.1 Construction of recombinant plasmids for A domain expression

To characterize the A domains *in* vitro, they should be isolated in enough amounts. To accomplish this, the A domains were heterologously expressed. For this the nucleotide sequences encoding the A domains were first identified by Pfam and BLAST. The identified gene of interest were PCR amplified from *H. aurantiacus* genomic DNA using proof reading *Pfu* DNA polymerase and primers with appropriate restriction sites for its insertion in to the pET-28a expression vector. To facilitate the expression of A domains, the PCR amplified product was cloned in to pET-28a expression vector.

The newly generated construct was stabilized by transforming in to a non-expression host, *E. coli* X-1 blue cells. Kanamycin resistance cassette, which is present in pET-28a vector, was used to screen the clones. The positive clones with the construct were identified using whole cell PCR. Moreover, the construct in the positive clones

was verified for any mutation by sequencing. This was followed by the transformation of the constructs isolated from the positive clones in to a suitable expression host.

3.5.6.2 Heterologous expression

In most part of this work *E.coli* based Rosetta (DE3) cells were used as the expression host. In special cases it is mentioned in results section. A single colony of the transformant was inoculated in 10 ml of LB medium containing 50 µg/ml of kanamycin as the selective marker. The culture was then shaked overnight at 37°C. On the next day, 1 L of LB medium containing 50 µg/ml of kanamycin was inoculated with 10 ml of the overnight culture. The culture was then shaked at 37°C till the culture reaches an OD_{600} of 0.5. After induction the culture was then shaked overnight at 16°C.

3.5.7 Recombinant protein purification and evaluation

3.5.7.1 Recombinant protein extraction

The cells were harvested from the induced cultures by centrifugation at 4000 rpm for 30 min at 4°C. The cell pellet was then resuspended in 5 ml of lysis buffer for native purification. 1 mg/ml of lysozyme was added to the suspension and incubated in ice for 30 min. The lysate was then sonicated for 6 X 10 sec each with a pause for 10 sec at 200-300 W. The sonication was carried out in ice. The insoluble materials were pelleted by centrifugation at 8000 rpm for 30 min at 4°C. The supernatant was the soluble fraction and contains proteins in its native form. The soluble and the insoluble fractions were analysed by SDS-PAGE for the presence of the protein of interest.

3.5.7.2 Recombinant protein purification

The six consecutive histidine residues (6X His-tag) conveyed by pET28-a vector at N- and C-terminus of the over-expressed protein facilitated the protein purification using Ni-NTA resin. The 6X His tag acts as chelator and immobilizes the protein on nickel ion in Ni-NTA resin.

1ml of Ni-NTA slurry (50% suspension) was added to the 1 ml polypropylene columns. The resin was washed with 5 CV of water to get rid of ethanol in the Ni-NTA

suspension. The column was equilibrated with 10 CV of lysis buffer. The proteins were always kept in ice. The soluble fraction was then added to the column and allowed to flow through. The process was repeated for at least two times. The fraction was then collected as flow through (FT). The column was then washed with 10 CV of wash buffer with 20mM imidazole to get rid of any non-specific binding. The fraction was collected as W1. The column was washed again with 10 CV of wash buffer with 40mM Imidazole to ensure that the column is free of non-specific binding contaminants. The fraction was collected as W2. The protein was then eluted from the column with 4X1 ml of elution buffer. The fractions were then analysed by SDS-PAGE. The eluates which are pure enough were then pooled together.

3.5.7.3 Buffer exchange and concentration of recombinant proteins

Imidazole interferes in further assay with the protein and hence the protein solution in imidazole was exchanged in A domain buffer using gravity-flow PD-10 desalting columns by GE Healthcare following the manufacturer's protocol. Since the amount of expressed proteins were low, concentration of the protein solutions in A domain buffer was done using Amicon Millipore columns (10KDa, 30KDa, 100KDa-depending on the molecular weight of the protein) as per the instructions from the manufacturer.

3.5.7.4 Polyacrylamide Gel Electrophoresis (PAGE)

The overexpressed protein of interest from the lysate and purified eluates were identified based on their molecular weight by sodium dodecyl sulfate (SDS)-PAGE (Laemmli 1970). The protein fraction to be analyzed was mixed with sample buffer in a ration of 3:1 and boiled at 100°C for 5 min. Depending on the molecular weight of the desired protein, 15 µl of protein samples were loaded on to 10% or 12% SDS-PAGE containing 1X Tris-glycine-SDS running buffer. To estimate the size of the protein, 10µl of unstained protein molecular weight standard was also loaded on to the SDS-PAGE. The electrophoretic run was performed in XCell SureLock MiniCell (Invitrogen Lifescience Technologies, Karlsruhe, Germany) at a constant voltage of 150V as per the maufacturer's instruction until the dye front reached the bottom of the gel. To visualize the separated proteins, the apparatus was disassembled and the gel was soaked in Coomassie staining solution for 30 min. Subsequently the

stained gel was soaked in destaining solution until the protein bands become visible and clear.

(A) Compounds for 4% stacking gel

0.5M Tris-HCl pH 6.0	1,250 µl
Bis-acrylamide (30%)	675 µl
10% SDS	25 µl
Water	3,050 µl
10% APS	25 µl
TEMED	2.5 µl

(B) Compounds for 12% separating gel

1M Tris-HCl pH 8.9	2,500 µl
Bis-acrylamide (30%)	675 µl
10% SDS	100 µl
Water	3,050 µl
10% APS	50 µl
TEMED	2.5 µl

Table 3-19: Composition of SDS-PAGE

The proteins for the MALDI-TOF-TOF measurements were separated in a native-PAGE. Native or denaturing PAGE was carried out in the absence of SDS in gels, running and sample loading buffers.

A) Compounds for 4% stacking gel

0.5M Tris-HCl pH 6.0	1,010 µl
Bis-acrylamide (30%)	404 µl
Water	2,585 µl
10% APS	40.4 µl
TEMED	4.04 µl

(B) Compounds for 10% separating gel

1M Tris-HCl pH 8.9	1,470 µl
Bis-acrylamide (30%)	1.951
Water	2,580 µl
10% APS	21 µl
TEMED	4.83 µl

(C) Sample buffer

0.5 M Tris- HCl pH 6.0	217 µl
Glycerol	508 µl
Water	275 µl
Bromophenolblue	50 µl

Table 3-20: Composition of native-PAGE

3.5.8 Biochemical analysis of recombinant proteins

3.5.8.1 γ-^{18}O$_4$-ATP pyrophosphate exchange assay

Adenylation proteins catalyze reversible activation of specific amino acids in the presence of ATP to form aminoacyl-adenylate and pyrophosphate (PPi). Conventionally this adenylation activity is monitored by a pyrophosphate exchange assay. Recently, a non-radioactive, mass spectrometry-based γ-^{18}O$_4$-ATP pyrophosphate exchange assay was developed in which the adenylation reaction was measured in reverse through the incorporation of γ-^{16}O$_4$-ATP from the isotope labeled γ-^{18}O$_4$-ATP in the presence of excess PPi (Phelan et al., 2009). The rate of mass shifts in the presence of different amino acids enables to conclude the specific amino acid that could be activated by the adenylation proteins *in vivo*.

The assay was carried out at in a total volume of 6 µl with 2µl of purified and concentrated protein, 2 µl of 3 mM amino acid containing 15 mM PPi in 20 mM Tris (pH-7.5) and 2µl of 3 mM γ-^{18}O$_4$-ATP containing 15mM MgCl$_2$ in 20mM Tris (pH-7.5). The assay mixture was incubated for 2 hrs at room temperature. The reaction was then quenched by adding 6 µl of 9-aminoacridine. 1 µl of this reaction mixture was analyzed for the mass shifts using MALDI-TOF MS. MALDI-TOF MS experiments were performed by Dr. Max crüsemaan in co-operation with the mass department from the Organic chemistry Institute, Bonn.

Data analysis

The quantification of the adenylation reaction was determined by the shift in mass from γ-^{18}O$_4$ isotope labelled ATP (m/z=514) to ^{16}O$_4$–ATP (m/z=506) using MALDI-TOF MS. Moreover, the masses of their monosodium adducts m/z=536 and m/z=528, respectively were also identified. In addition to the exchange into ^{16}O$_4$–ATP, other masses were also detected corresponding to the incomplete exchange ^{18}O$_3$–ATP (m/z=512), ^{18}O$_2$–ATP (m/z=510), ^{18}O$_1$–ATP (m/z=508) and their corresponding sodium adducts m/z=534, m/z=532 and m/z=530, respectively. Percentage of exchange was calculated by the ratio of area of unlabelled ATP (m/z=506) to the sum of the area of all the ATP species including unlabelled (m/z=506), completely (m/z=514) and partially labeled (m/z=508, m/z=510 and

m/z=512) and their monosodium adducts (m/z=528, m/z=530, m/z=532, m/z=534 and m/z=536).

3.5.8.2 Phosphopantetheinylation of peptidyl carrier protein (PCP)

The building blocks of nonribosomal peptides are attached to the phosphopantetheine arm of PCP in the NRPS. This phosphopantetheine arm on the PCP is derived from coenzyme A by the action of phosphopantetheinyl transferase (PPTase). This phosphopantetheinylation, which converts PCP from its inactive apo to holo form is crucial for the functioning of NRPS and can be identified *in vitro* by shift in the mass of 340.1 Da due to the phosphopantetheine arm.

Heterologously expressed A-PCP didomain (1mg/ml) and PPTase (0.4 mg/ml) was incubated with 8 mM $MgCl_2$ and 320 µM CoA for 1 hr at room temperature. 100 µl of the so formed holo- didomain was incubated with 5 µl of 0.1 M ATP and 1 mM of amino acid for the loading of specific amino acid activated by the A domain on to the PCP domain. The reaction was carried out for 30 min at room temperature before stopping it by adding equal volume of 10% formic acid. The proteins were then separated via native-PAGE for MALDI-TOF/TOF analysis.

4. Results

4.1 Mining the genomes of gliding bacteria

There are more than thirty accessible gliding bacterial genomes in NCBI, the majority of which belong to the phylum Bacteroidetes. Published cyanobacterial genomes were excessively analyzed for the PKS and NRPS genes (Kalaitzis et al., 2007; Donadio et al., 2007) and are thus not further considered. The focus here is instead put on other gliding bacteria belonging to the phylum Proteobacteria, Planctomycetes, Bacteroidetes and Chloroflexi. Genomes were analyzed for PKS and NRPS genes computationally using CLUSEAN (Weber et al., 2009), BLAST, Pfam, PKS/NRPS Analysis (Bachmann and Ravel, 2009), NRPS predictor (Rausch et al., 2005) and NRPS/PKS database (Ansari et al., 2004). The genome of *H. aurantiacus* DSM 785 from the phylum Chloroflexi is analysed in detail in order to predict the putative structures encoded by PKS and NRPS genes of their genome.

4.1.1 Genomes of the phylum Proteobacteria

This phylum includes the myxobacteria from which compounds with unprecedented structural and biological properties were reported (Wenzel et al., 2005). The genome sequences of *Myxococcus xanthus* DK 1622, *Myxococcus fulvus* HW-1, *Sorangium cellulosum* So ce 56 and *Stigmatella aurantiaca* DW4/3-1, three *Anaeromyxobacter* species (*Anaeromyxobacter dehalogenans* 2CP-1, *Anaeromyxobacter sp. K, Anaeromyxobacter sp.* FW 109-5) and of *Haliangium ochraceum* DSM 14365 were reported and allow to gain further insight into the potential metabolic profile of the respective bacterial strains. The genomes of *Anaeromyxobacter* species comprise approximately 5 Mb and were reported to be devoid of PKS and NRPS genes, and indeed so far no secondary metabolites are reported from the latter (Wenzel et al., 2009). For *So. cellulosum* Soce56, *M. xanthus* DK1622 and *Sg. aurantiaca* DW4/3-1 biosynthetic genes for specific natural products were already annotated and were thus not considered further on (Bode and Müller, 2006; Li and Müller, 2009; Wenzel and Müller, 2009b). The genomes of *M. fulvus* HW-1 (Li et al., 2011) and *Ha. ochraceum* DSM 14365 (Ivanova et al., 2010) which were sequenced completely and published recently are here analysed for the presence of PKS and NRPS genes.

4.1.1.1 *Myxococcus fulvus* HW-1

M. fulvus HW-1 is a halotolerant myxobacterial strain isolated from a coastal sea-water sample. The strain has a circular genome of 9 Mb in size (Li et al., 2011). Although potent antibiotics like myxopyronins, myxathiazol, myxovalargins and cytotoxic bithiazole metabolites are known for *M. fulvus* strains, none of these metabolites have to date been isolated from *M. fulvus* HW-1. In contrast, scanning the genome revealed a pool of PKS and NRPS genes, which however did not correspond to the above mentioned structural classes.

Just like in other myxobacterial genomes, the PKS/NRPS hybrid gene loci are predominant over the discrete NRPS and PKS gene loci. From the genome sequence, nine PKS/NRPS hybrid gene loci were identified, whereas only six NRPS and two PKS gene loci were recognized (Table 4-1). Two of the gene loci could be attributed to known structural classes, i.e. the Dxanthenes and myxochelin A. Thus, the PKS/NRPS gene locus-8 putatively encodes the biosynthetic enzymes for DKxanthenes (Figure 4-1, Table 4-1). It has the approximate size of 47.5 Kb including, as expected seven PKS and two NRPS modules. Till date, none of the DKxanthenes were yet isolated from the strain.

(A)

M. xanthus **DK1622**

M. fulvus **HW-1**

(B)
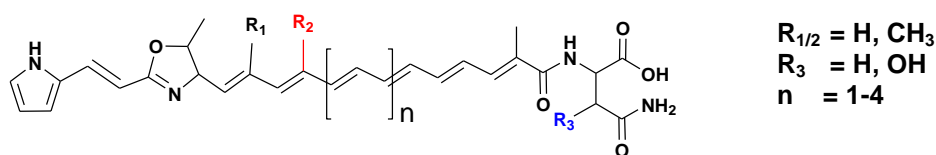
Figure 4-1: DKxanthenes in *M. fulvus* HW-1.

(A) The DKxanthene biosynthetic gene loci from *M. xanthus* DK1622 and *M. fulvus* HW-1. Red, PKS genes; blue, NRPS genes; green, PKS/NRPS hybrid genes; grey, accessory genes; black, a hypothetical protein in different locations; pink, starter unit biosynthetic genes; orange, thioesterase.

(B) Structure of DKxanthenes.

The NRPS gene locus-5 is highly homologous to the myxochelin A biosynthetic gene cluster (Gaitatzis et al., 2001; Silakowski et al., 2000). The locus includes the genes required for the synthesis of 2,3-dihydroxybenzoic acid (Kwon et al., 1996) and for iron transport and utilization (Silakowski et al., 2000). The respective A domain has the Stachelhaus code, DAEDIGTVVK, which is consistent with that of the A domain in myxochelin A. Therefore, the NRPS gene locus-5 in *M. fulvus* HW-1 is proposed to encode for the biosynthesis of myxochelin A.

(A) *M. xanthus* **DK1622**

M. fulvus **HW-1**

(B)
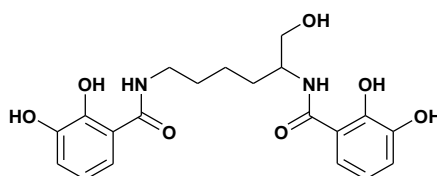
Figure 4-2: Myxochelin in *M. fulvus* HW-1.

(A) Myxochelin biosynthetic gene loci from *Myxococcus xanthus* DK1622 and *Myxococcus fulvus* HW-1. Red, NRPS gene; green, aminotransferase; grey, iron transport and utilization proteins, others, genes for the biosynthesis and loading of DHBA.

(B) Structure of myxochelin A.

Gene	Gene locus[a]	Size[b] (Kb)	No. of modules		Compound
			PKS	**NRPS**	
PKS/NRPS-1	*lilab_9160-lilab_9195*	21.4	1+ACL	2	Unknown
PKS/NRPS-2	*lilab_21685-lilab_21695*	11.8	1	1	Unknown
PKS/NRPS-3	*lilab_26120-lilab_26145*	30.3	3	5	Unknown
PKS/NRPS-4	*lilab_26215-lilab_26265*	71.8	1	19	Unknown
PKS/NRPS-5	*lilab_27065-lilab_27075*	27.6	1	13	Unknown
PKS/NRPS-6	*lilab_28140-lilab_28150*	28.8	1	8	Unknown
PKS/NRPS-7	*lilab_ 28580 -lilab_28585*	22.6	1	6	Unknown

PKS/NRPS-8	lilab_29600-lilab_29675	47.5	7	2	DKxanthenes
PKS/NRPS-9	lilab_30100-lilab_30125	33.2	1	7	Unknown
NRPS-1	lilab_0695-lilab_0700	17	-	5	Unknown
NRPS-2	lilab_0720	4	-	1	Unknown
NRPS-3	lilab_1140	4.4	-	1	Unknown
NRPS-4	lilab_24940-lilab_24945	13.2	-	4	Unknown
NRPS-5	lilab_26275-lilab_26330	20.2	-	1	Myxochelin A
NRPS-6	lilab_30650	6.5	-	2	Unknown
PKS-1	lilab_25050-lilab_25055	8	1	-	Unknown
PKS-2	lilab_26180-lilab_26200	10.9	2	-	Unknown

Table 4-1: PKS and/or NRPS gene loci from *M. fulvus* HW-1. [a] The gene loci are named according to GenBank, [b] Unannotated nucleotide sequences between each gene locus are also considered. ACL- acyl-CoA synthetase.

4.1.1.2 *Haliangium ochraceum* DSM 14365

Recently the 9.4 Mb genome of *H. ochraceum* DSM 14365 was completely sequenced and published (Ivanova et al., 2010). This was the first obligate marine myxobacterium to have its genome sequenced. Our scanning of the genome sequence revealed three PKS/NRPS hybrid gene loci, eight NRPS, and three PKS loci (Table 4-2). So far, none of these gene clusters can be attributed with certainty to a known natural product. However, there are some aspects that let us consider PKS gene locus-3 as the gene cluster encoding haliangicin biosynthetic enzymes. Haliangicin an antifungal compound is the only secondary metabolite isolated from *Ha. ochraceum* DSM 14365 (Fudou et al., 2001; Kundim et al., 2003).

Putative haliangicin biosynthetic gene cluster

Haliangicin is a polyunsaturated metabolite with an epoxide ring, a β-methoxy acrylate at one and an alkene moiety at the other end of the molecule (Figure 4-3A) [Fudou et al., 2001). Haliangicin occurs as *cis* and *trans* isomer at the epoxide portion of the molecule. In addition, there are geometrical isomers at the polyene moieties i.e. haliangicin B-D, which were also isolated as *cis, trans* mixtures (Figure

4-3-A) (Kundim et al., 2003). Concluded from the structure, the biosynthesis requires seven PKS elongation steps followed by post-PKS modification reactions like *O*-methylation, epoxide ring and alkene formation. The methoxy groups at the terminal carboxyl group C1 and at the β-carbon (C3) are likely to be added by an *O*-MT. On the contrary, the methoxy substitution at C4 is most likely to be derived from the building block methoxy malonate. A striking feature is the methyl branching at C9. This type of β-methyl branching is incorporated by the action of HMG-CoA synthases (Gulder et al. 2011) and ECH enzymes. Considering these factors the genome was scanned for such PKS genes. PKS gene locus-3 has all the genes (Figure 4-3B) required for the synthesis of methoxy malonyl-CoA (Chan et al., 2006) (Figure 4-4) and the incorporation of β-methyl branching (Figure 4-5). Since the locus has merely six elongation modules, one of the PKS modules is most likely used iteratively. Since the biosynthetic genes are not strictly co-linear, the elucidation of the exact biosynthetic pathway requires feeding and gene inactivation studies. Especially the formation of terminal alkene would be an interesting feature to study further.

(A)

Compound	Geometry			
	C2	C4	C6	Epoxide[a]
Haliangicin	*E*	*Z*	*Z*	*trans*
cis-Haliangicin	*E*	*Z*	*Z*	*cis*
Haliangicin B (+ *cis* isomer)	*Z*[b]	*Z*	*Z*	*trans* (+*cis*)
Haliangicin C (+ *cis* isomer)	*E*	*E*[b]	*Z*	*trans* (+*cis*)
Haliangicin D (+ *cis* isomer)	*E*	*E*[b]	*E*[b]	*trans* (+*cis*)

(B)

Hoch_1169-Hoch_1187
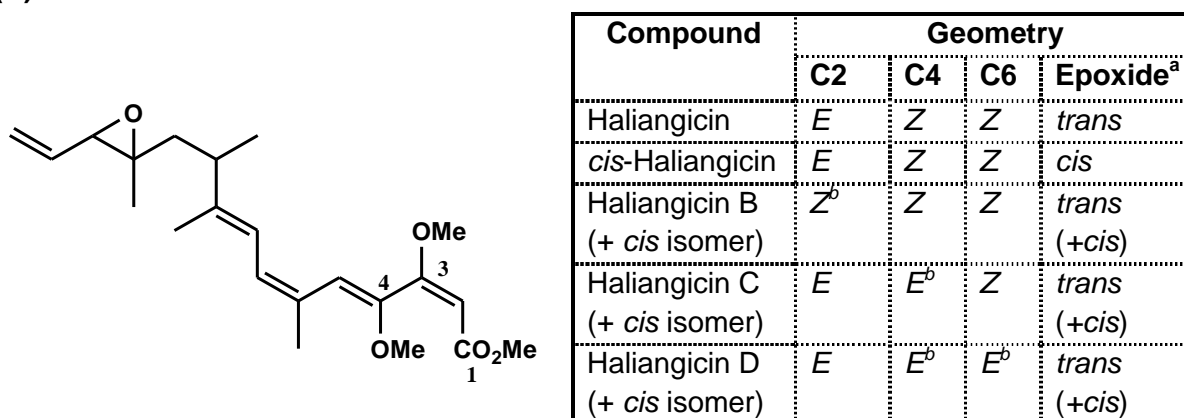
Figure 4-3: Haliangicin in *Ha. Ochaceum* DSM 14365.

(A) Structure of haliangicin. The table was adapted from Kundim et al and lists the isomers of haliangicin. [a] *cis* isomers were the minor components.

(B) Proposed haliangicin biosynthetic gene cluster. PKS genes are shown in red. Genes involved in the synthesis of methoxymalonyl-CoA are shown in green. Putative genes involved in the HMG-CoA synthase reaction are shown in blue. However, a discrete gene

encoding KS domain for the HMG-CoA synthase reaction is missing. MT gene is shown in violet. Other accessory genes are shown in grey.
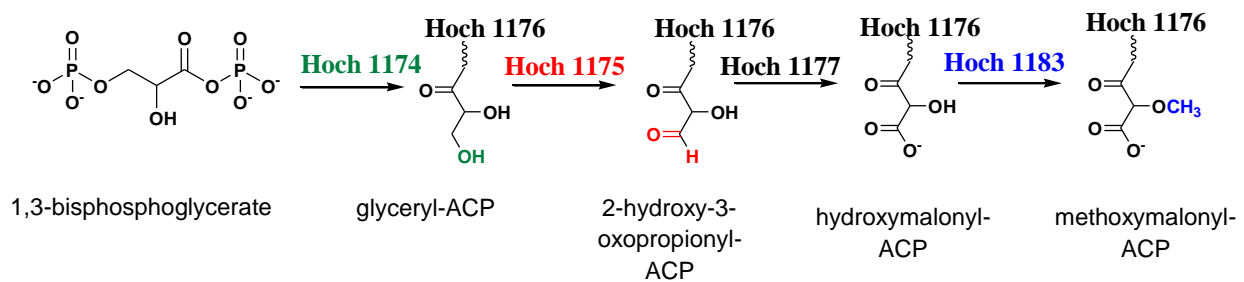
Figure 4-4: Biosynthesis of methoxymalonyl-CoA. Hoch 1174, glyceryl transferase (phosphatase); Hoch 1176, acyl carrier protein (ACP); Hoch 1175, acyl-CoA dehydrogenase; Hoch 1177, β-hydroxyacyl CoA dehydrogenase; Hoch 1183, *O*- methyltransferase.
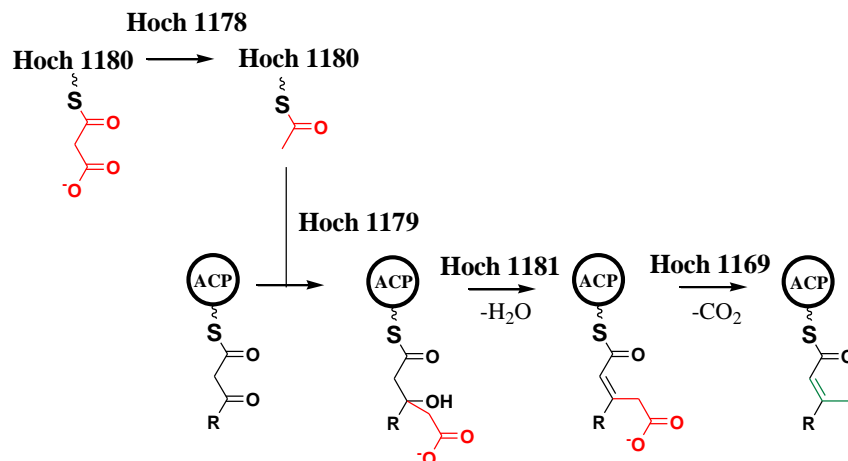
Figure 4-5: Beta branching mediated by Hoch 1180 (ACP), Hoch 1178 (KS), Hoch 1179 (HMG-CoA synthase), Hoch 1181 (ECH) and Hoch 1169 (ECH). ACP, Acyl Carrier Protein; KS, Ketosynthase; HMG-CoA synthase, Hydroxymethyl glutaryl-Coenzyme A synthase; ECH, Enoyl-Coenzyme A hydratase.

Siderophore-type metabolite

Scanning of the genome sequence exposed the NRPS gene locus-4, which is bordered by genes required for the synthesis of 2,3-dihydroxybenzoic acid (DHBA) (Kwon et al., 1996), for iron transport and utilization. Thus, the assumption was made that NRPS gene locus-4 encodes siderophore biosynthetic enzymes. The biosynthetic pathway resembles that of myxochelin A (Figure 4-2) (Gaitatzis et al., 2001; Silakowski et al., 2000). Based on the Stachelhaus code, the A domain was

predicted to activate lysine as for myxochelins. But, in contrast to the myxochelin A cluster, the gene locus includes a methyltransferase gene, *hoch_2016*. Therefore, it is predicted that the NRPS gene locus-4 of *Ha. ochraceum* DSM 14365 encodes for a methylated derivative of myxochelin A. Clearly, the isolation of the putative siderophore and structure elucidation is necessary to gain further proof.

Gene	Gene locus	Size (Kb)	No. of modules		Compound/Comments
			PKS	**NRPS**	
PKS/NRPS-1	*hoch_0798-hoch_0799*	21.7	4	1	Unknown
PKS/NRPS-2	*hoch_1746-hoch_1750*	40	2	8	Unknown
PKS/NRPS-3	*hoch_2945-hoch_2957*	59	8	2	Unknown
NRPS-1	*hoch_1077*	12.3	-	4	Unknown
NRPS-2	*hoch_1312*	3.7	-	1	Unknown
NRPS-3	*hoch_1365-hoch_1371*	25.9	-	6	Unknown
NRPS-4	*hoch_2016-hoch_2022*	11.4	-	1	Methylated myxochelin A (putative)
NRPS-5	*hoch_1694*	4.06	-	1	Unknown
NRPS-6	*hoch_5720*	1.5	-	1	Unnown
NRPS-7	*hoch_6170*	3.2	-	1	Unnown
NRPS-8	*hoch_6298*	2.7	-	1	Unknown
PKS-1	*hoch_0692-hoch_0693*	6.4	1	-	Unknown
PKS-2	*hoch_2966-hoch_2975*	52.6	9	-	Unknown
PKS-3	*hoch_1168-hoch_1187*	46.1	6	-	Haliangicin (putative)

Table 4-2: PKS and/or NRPS hybrid gene loci from *H. ochraceum* DSM 14365

4.1.2 Genome from the phylum Planctomycetes

Isospahaera is the only genus in the phylum Planctomycetes to possess gliding motility. The sequences of the 5.5 Mb genome isolated from *Isosphaera pallida* ATCC 43644 (Göker et al., 2011) reveals only one PKS gene locus. This PKS gene locus from Isop 2002-Isop 2006 has a size of 17 Kb and includes genes encoding for

a 3-oxoacyl-ACP and a α/β hydrolase. The locus comprises a single PKS module split between two ORFs with tandem ACP domains.

4.1.3 Genomes from the phylum Bacteroidetes

With 12 genomes accessible in NCBI, the biosynthetic potential of Bacteroidetes seems not as prolific as that of the phylum Proteobacteria. Eight of the genomes were less than 5 Mb in size and thus they were unlikely to possess NRPS and PKS biosynthetic gene clusters. *Spirosoma lingulae* DSM 74 with a genome of around 8.1 MB (Lail et al., 2010) could be expected to have many PKS and NRPS genes. However, no genes for PKS or NRPS metabolism could be identified from this genome. The 6.3 Mb genome of *Cytophaga fermentans* DSM 9555 was also lacking PKS and NRPS genes. The 6.1 Mb genome sequence of *Flavobacterium johnsoniae* UW-101 was already reported to harbor an NRPS gene cluster from Fjoh 2083-Fjoh 2104 spanning approximately 70.7 Kb (McBride et al., 2009). However, the associated nonribosomal peptide product is not yet isolated. PKS and NRPS genes were identified from *Chitinophaga pinensis* DSM 2588 genome sequences (Rio et al., 2010). Furthermore, the genome of *Paedobacter heparinus* DSM 2366 was reported (Han et al., 2009) and we analyzed for the PKS and NRPS biosynthetic genes.

4.1.3.1 *Chitinophaga pinensis* DSM 2588

Chitinophaga pinensis DSM 2588 is the first organism to be completely sequenced in the family Chitinophagacae. The organism harbors a genome of 9.1 Mb in size (Rio et al., 2010). A *trans* AT-PKS gene cluster of 75.1 Kb had already been reported to encode for the biosynthesis of elansolid B_1 and D (Figure 4-6) (Teta et al., 2010). In this context, it is interesting to note that Müller et al. described the isolation of elansolids A_1/A_2, elansolid A_3 and elansolids B_1-B_3 from a different *Chitinophaga* sp., i.e. *C. sanctii* (Dehn et al., 2011). From the *C. pinensis* DSM 2588 genome NRPS genes in the size of 98 Kb spanning from Cpin_5188- Cpin_5195 and a PKS/NRPS mixed genes in the size of 88 Kb spanning from Cpin_3399- Cpin_3421 were identified during the current study. These NRPS and PKS/NRPS biosynthetic gene loci could not be correlated with any known compounds.

Elansolid B₁   **Elansolid D**
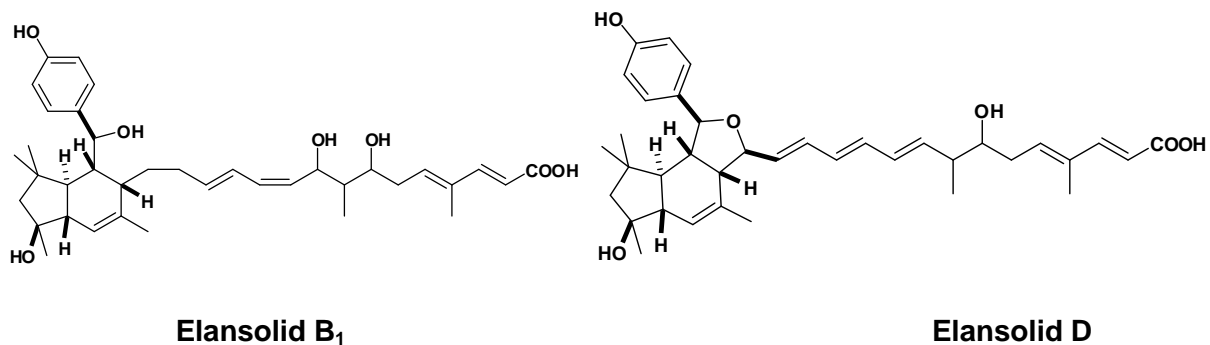
Figure 4-6: Structures of elansolids isolated from *C. pinensis* DSM 2588.

4.1.3.2 *Paedobacter heparinus* DSM 2366

No metabolites have to date been isolated from members of this genus. The genome sequence of *Paedobacter heparinus* DSM 2366 (Han et al., 2009) harbors only one PKS/NRPS hybrid gene locus with a size of 12.5 Kb spanning from Phep 2032- Phe 2034. The locus includes two NRPS and a single PKS module. As in the case of the myxothiazol A biosynthetic gene cluster (Wenzel and Müller, 2009b), downstream to the NRPS and PKS genes is the 4'-phosphopantetheinyl transferase gene. The respective gene product is necessary for the priming the biosynthetic enzymes.

4.1.4 Genomes from the phylum Chloroflexi

Till date, only four genomes from this phylum were sequenced completely. Out of these, three belong to the genus *Chloroflexus (Ch. aurantiacus J-10-fl, Chloroflexus sp. Y-400-fl, Ch. aggregans DSM 9485)* with the genome size of 5.3 Mb or smaller. Moreover, we could not identify PKS or NRPS genes in *Chloroflexus* genomes. The fourth completed genome sequence is that of the bacterium *Herpetosiphon aurantiacus* DSM 785.

4.1.4.1 *Herpetosiphon aurantiacus* DSM 785

The genome of *H. aurantiacus* DSM 785 harbors two plasmids of 3.4 Kb and 1 Kb size, respectively and a single circular chromosome of 6.3 Mb. This circular chromosome is the largest among all members of the Chloroflexi. It was sequenced completely by the Joint Genome Institute, annotated and made publicly accessible in NCBI. Analysis for PKS and NRPS genes revealed five PKS/NRPS (Table 4-3)

hybrid, five NRPS and a single PKS gene locus from the circular chromosome. However, no PKS and NRPS genes were identified from the plasmids.

With 6.3% of the genome contributing to the synthesis of PKS- and NRPS-derived compounds, the biosynthetic ability of this bacterium is comparable to that of other and well known secondary metabolite producers such as *Streptomyces coelicolor* (4.5%), *S. avermitis* (6.6%) and *Pseudomonas fluorescens* (5.7%) (Nett et al., 2009). But this sequenced bacterial strain is not known for the production of secondary metabolites yet, raising the question whether the gene clusters present are indeed expressed. In a related strain, however the compound siphonazole was obtained as the first secondary metabolite from bacteria of the genus *Herpetosiphon* (Nett et al., 2006). No gene cluster related to the biosynthesis of siphonazole was identified from the *H. aurantiacus* DSM 785 genome sequence.

Gene	Gene locus	Size (Kb)	No. of modules		Compound
			PKS	**NRPS**	
PKS/NRPS-1	*haur_1856-haur_1863*	53.1	2+ACL	6	Unknown
PKS/NRPS-2	*haur_1871-haur_1888*	71.9	1	11	Unknown
PKS/NRPS-3	*haur_2008*	12.3	1	1	Unknown
PKS/NRPS-4	*haur_2406-haur_2419*	35.9	1	3	Unknown
PKS/NRPS-5	*haur_3958-haur_3974*	112.6	16	8	Unknown
NRPS-1	*haur_1571-haur_1574*	10	-	3	Unknown
NRPS-2	*haur_1800-haur_1809*	14	-	1	Myxochelin A
NRPS-3	*haur_2088-haur_2092*	16	-	5	Unknown
NRPS-4	*haur_2104-haur_2109*	18.5	-	4+ACL	Unknown
NRPS-5	*haur_3124-3130*	13.3	-	2	Unknown
PKS-1	*haur_0864-haur_0874*	17.2	1	-	Unknown

Table 4-3: PKS and/or NRPS gene loci from *H. aurantiacus* DSM 785. ACL, acyl-CoA synthetase

Therefore, to explore the biosynthetic potential of *H. aurantiacus* DSM 785, the sequences of the five PKS/NRPS, five NRPS and PKS gene loci are explored here in detail to gain perspectives towards the strategies to isolate the respective compounds.

4.1.4.1.1 PKS/NRPS gene locus-1

The hybrid gene locus-1, comprising six NRPS and two PKS modules is organized in five ORFs, i.e *haur_1857, haur_1858, haur_1859, haur_1860 and haur_1861* (Table 4-4). One of the PKS modules was found to be the PKS/NRPS hybrid module-1 (Haur_1857). All PKS and NRPS ORFs have a size of 34.6 Kb and are flanked by additional ORFs speculated to be involved in the biosynthesis and modification of the PKS/NRPS hybrid product. These are transcriptional regulators (Haur_1853 and Haur_1855) and an AMP-dependant synthetase and ligase (Haur_1856) in the upstream and a transporter (Haur_1862), a pyridoxal dependent (PLP) enzyme (Haur_1863) as well as ornithine cyclodeaminase (Haur_1864) in the downstream region (Table 4-4).

Gene locus	Modules	Putative functions
haur_1856		AMP-dependant synthetase and ligase
haur_1857	Module-L Module-1 Module-2	ACP_L KS_1-AT_1-PCP_1-AMT_1 C_2-A_2-MT_2-PCP_2
haur_1858	Module-3	C_3-A_3-PCP_3-E_3
haur_1859	Module-4 Module-5	C_4-A_4-PCP_4-E_4 C_5-A_5-PCP_5
haur_1860	Module-6	KS_6-AT_6-KR_6-ACP_6
haur_1861	Module-7 Module-8	C_7-A_7-PCP_7-E_7 C_8-A_8-PCP_8-TE
haur_1862		Pyridoxal-phosphate dependent enzyme
haur_1863		Ornithine cyclodeaminase

Table 4-4: List of genes in PKS/NRPS gene locus-1 and their putative functions. Suffices indicate the respective modules. L, Loading module; ACP, Acyl Carrier Protein; KS,

Ketosynthase; AT, Acyltransferase; AMT, Aminotransferase; C, Condesation; A, Adenylation; MT, Methyltransferase; PCP, Peptidyl Carrier Protein; E- Epimerization

AMP-dependent synthetase and ligase

AMP-dependent synthetases and ligases belong to the adenylate-forming family of enzymes (ANL), which includes acyl and aryl-CoA synthetases, A domains of NRPS and luciferase (Gulick 2009). ANL enzymes in general catalyze the activation of a carboxylate substrate at the expense of ATP to form an acyl adenylate intermediate. The ANL enzymes can be easily identified from the amino acid sequences based on the three conserved core motifs, i.e. motif-1 (A3)-L-A-Y-X-X-Y-T-S-G-S(T)-T-G-X-P-K-G, motif-2 (A5)- Y-G(W)-P(X)-T(A)-E and motif-3 (A7)- Y(X)-R(K)-T(S)-G-D-L, all required for ATP binding (Gulick 2009). The alignment of Haur_1856 AMP dependent synthetase and ligase showed 48% identity to the AMP-forming acyl-CoA synthetases of *Moorea producta* 3L. Also, the Haur_1856 sequences contain the three conserved motifs for ATP binding (Figure 4-7). In addition, 34 residues downstream to motif I is a histidine residue typical for long chain acyl-CoA synthetases and luciferase. In the case of A domains, this typical amino acid is phenylalanine and in small chain acyl-CoA synthetases it is tryptophan (Gulick 2009). Thus, it is concluded that based on bioinformatic results AMP-dependent synthetase and ligase from PKS/NRPS gene locus-1 is adenylating a long chain carboxy acid.

| | L | A | Y | X | X | Y | T | S | G | S | T | G | X | | P | K | G | | | H | | | Y | G | P | T | E | | | Y | R | T | G | D | L | |
											T														W	X	A				X	K	S			
Haur 1856	L	A	F	L	Q	Y	T	S	G	S	T	S	Q		P	R	G	193		H	228		Y	G	L	A	E	339		L	R	T	G	D	L	448
M. producta	L	A	F	L	Q	Y	T	S	G	S	T	A	A		P	K	G	181		H	216		Y	G	M	A	E	328		L	R	T	G	D	L	441
HctA	I	A	Y	L	Q	Y	T	S	G	S	T	S	T		P	K	G	185		H	220		Y	G	L	A	E	332		L	R	T	G	D	L	445
Beta-ketoacyl	L	A	F	L	Q	Y	T	S	G	S	T	G	T		P	K	G	187		H	222		Y	G	M	A	E	335		L	R	T	G	D	L	447
LC-FA-CoA	L	A	F	L	Q	Y	T	S	G	S	T	G	N		P	K	G	175		H	210		Y	G	L	A	E	323		L	R	T	G	D	L	446

Figure 4-7: Characteristic motifs in acyl-CoA synthetases (highlighted in yellow). Alignment of AMP-dependent synthetase and ligase from PKS/NRPS gene locus-1 (Haur 1856) of *Moorea producta* (Acc no-ZP08428587.1), of HctA from *Lyngbya majuscula* (Acc no-AAY42393.1), of β-keto acyl synthase-AMP from *Microcoleus chthonoplastes* PCC 7420 (Acc no- ZP05029386.1) and Long chain (LC) Fatty acid (FA)-CoA ligase from *Cyanotheca* sp (Acc no- ZP08976119.1). The histidine amino acid typical for acyl-CoA synthetases selecting long chain substrates is highlighted in orange. The consensus sequences from the literature are highlighted in bold. The numbers indicate the position of amino acids.

Adenylation domains

Non-ribosomal peptides possess diverse structures because of the multiple building blocks. These comprise not only the twenty proteinogenic amino acids but also other acids (Schwarzer et al., 2003). It is the A domain that accesses the specific substrate for loading to the respective NRPS module (Marahiel et al., 1997). The A domain amino acid sequences contain ten highly conserved core motifs (Konz and Marahiel, 1999). These core motifs harbor the ten amino acid residues referred to as Stachelhaus code. They are found in between the core motifs-4 and -5, and in motif-10, and together define the substrate of the respective A domain (Stachelhaus et al., 1999, Challis et al., 2000).

The hybrid gene locus-1 with six NRPS modules comprises six A domains, i.e A_2–A_8 (Table 4-4). All the A domains contain the ten core motifs with a slight deviation (Figure 4-8A). The ten amino acid code extracted from in between the core motifs-4, 5 and -10 were used to predict the substrate for each A domain. The Stachelhaus code, DIWELTADD deduced from the A_2 domain amino acid sequence is 90% identical to that of the A domain predicted to activate 2,4-diaminobutyric acid in pyoverdine biosynthesis (Mossialos et al., 2002). The computational analysis of the A_2 domain amino acid sequences in the NRPS predictor-2 showed a high score for the amino acid 4-hydroxyphenlyglycine (HPG) and its derivatives. Despite these results, the putative substrate activated by the A_2 domain remains obscure. Except for A_2, the substrate specificities of all A domains could be predicted using the Stachelhaus code (Figure 4-8B). Thus, A domain A_3 activates asparagine, A_4 activates threonine, A_5 activates serine, A_7 activates ornithine and A_8 activates glycine.
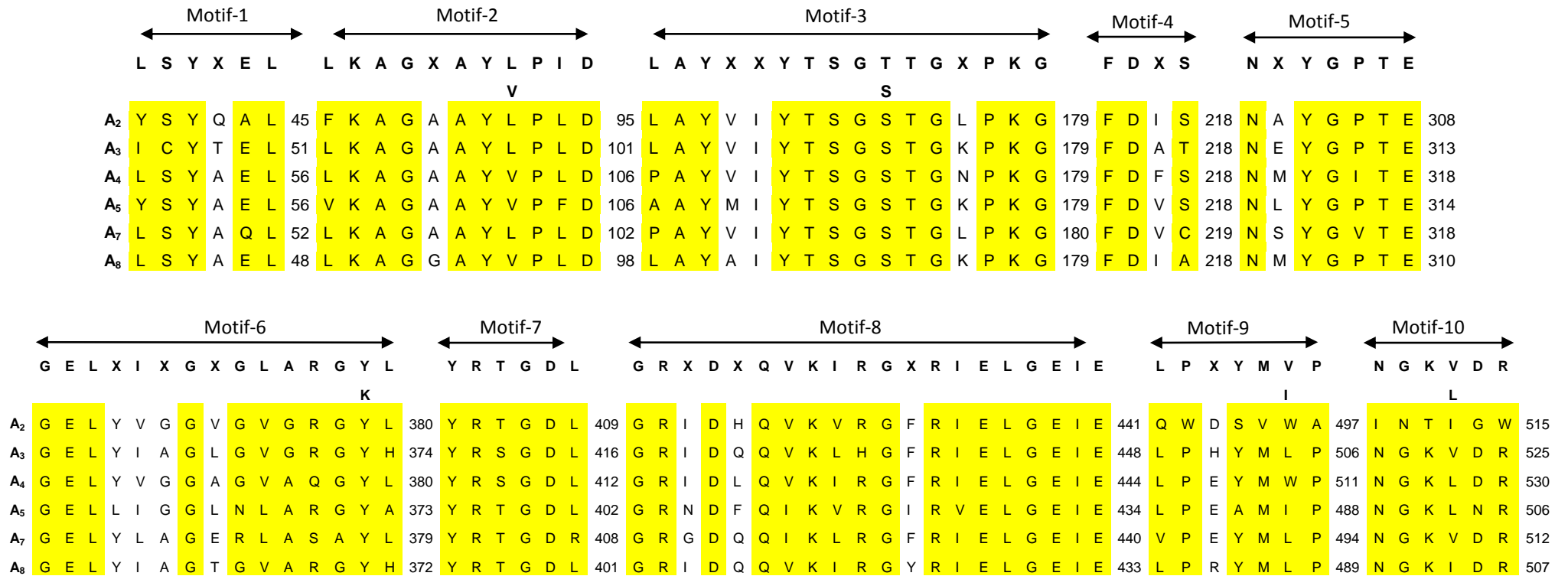
(A)

(B)

Gene locus	A domain	Stachelhaus code	Predicted substrate
haur_1857	A_2	DIWELTADD*	Unknown
haur_1858	A_3	DATKVGEVGK	Asparagine
haur_1859	A_4	DFWNIGMVHK	Threonine
haur_1859	A_5	DVWHFSLVDK	Serine
haur_1861	A_7	DVGEIGSIDK	Ornithine
haur_1861	A_8	DILQLGMVWK	Glycine

73

Figure 4-8: PKS/NRPS gene locus-1 adenylation (A) domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS A domains from PKS/NRPS gene locus-1. The ten core motif sequences (Konz and Marahiel, 1999) are highlighted in bold and the respective motifs in the PKS/NRPS-1 A domains are highlighted in yellow. The numbers represent the amino acid positions in the respective proteins. Suffices indicate the respective modules.

(B) The table lists the Stachelhaus codes extracted from the sequences of PKS/NRPS gene locus-1 A domains and their putative amino acid substrates. The asterisk indicates the missing amino acid.

Acyltransferase domains

In PKS megasynthetases the AT domain is the key element in determining the building blocks of the respective polyketide. The active site motif-2 of the AT domains include the sequence GHSIG, where the serine is required for the catalytic activity of the AT. In the PKS/NRPS gene locus-1 two AT domains, i.e AT_1 and AT_6 (Table 4-4) are encoded, which both have this sequence motif (Figure 4-9). The two other conserved motifs, i.e. motif-1 and -3 that were identified in AT1 and AT6 domains showed slight variation in their sequence but showed close homology to malonyl-CoA specific ATs (Smith and Tsai, 2007)(Figure 4-9). Therefore, both the AT domains was speculated to select malonyl-CoA as their substrates.

	Motif-1							Motif-2						Motif-3				
	Z	T	X	$	A	Q		G	H	S	X	G		*	A	F	H	
					T	E								G		S		
AT1	D	T	R	Y	T	Q	62	G	H	S	I	G	92	H	G	F	H	191
AT6	A	T	N	V	A	Q	60	G	H	S	I	G	90	H	A	F	H	187
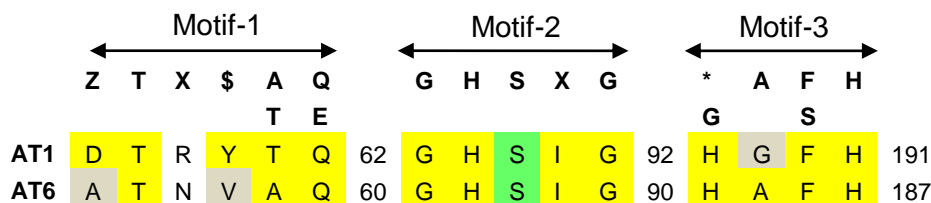
Figure 4-9: Characteristic motifs in the AT domains, AT_1 and AT_6 (highlighted in yellow) of PKS/NRPS gene locus-1. The catalytically active serine is highlighted in green. Malonyl-CoA specific AT consensus sequences from the literature (Smith and Tsai, 2007) are highlighted in bold. Z, hydrophilic amino acid; $, aromatic amino acid; X, branched hydrophobic amino acid; *, [HTVYGMIL]. The residues that deviates from the literature are highlighted in brown. The numbers represent the position of the amino acids. Suffices indicate the respective modules.

Carrier protein (ACP-PCP) domains

After the selection of the substrates by the respective AT or A domains, they are loaded to the 4'-phosphopantetheine arm attached to the serine residue of the carrier proteins, referred to as ACP in PKS and PCP in NRPS (Lai et al., 2006). Sequence alignment of the ACP and the PCP domains of the hybrid gene locus-1, i.e PCP_2 to PCP_8 and ACP_L, ACP_1 as well as ACP_6 (Table 4-4) showed that the active site serine residue (Findlow et al., 2003) is conserved, suggesting that all the carrier protein domains are active (Figure 4-10). The other conserved sequences were also identified (Konz and Marahiel, 1999; Pawlik et al., 2007).

(A) PCP domains   **(B) ACP domains**

Figure 4-10: Characteristic motifs in the PCP (A) and ACP (B) domains (highlighted in yellow) of PKS/NRPS gene locus-1. The PPant attaching serine residue is highlighted in green. Conserved sequences from the literature (Konz and Marahiel, 1999; Pawlik et al., 2007) are highlighted in yellow. The numbers indicate the position of amino acid. Suffices indicate the respective modules. L- Loading module.

Condensation domains

The C domains are responsible for connecting all the loaded substrates in a sequential manner by a condensation reaction. This takes place between the upstream amino acyl or peptidyl-S-PCP with the downstream aminoacyl-S-PCP, resulting in amide bonds (Schwarzer et al., 2003). Albeit the C domains in NRPS and KS domains in PKS have analogous functions, the chemical mechanisms behind the connecting reactions are divergent (Fischbach et al., 2006).

C domains include seven highly conserved core motifs (Konz and Marahiel, 1999). Alignment of the hybrid gene locus-1 C domains, i.e C_2 - C_8 (Table 4-4) showed that there were slight deviations in the conserved sequences in six of the seven core

motifs (Figure 4-11). Just like in the myxochelin biosynthetic gene cluster, the seventh motif could not be identified from the sequence (Silakowski et al., 2000). However, all the hybrid gene locus-1 C domains harbor the highly conserved H-H-X-X-X-D-G region. This histidine motif is necessary for amide bond formation (Bergendahl et al., 2002; Konz and Marahiel, 1999) and its presence makes it likely that all C domains of the hybrid gene locus-1 are functional. One of the C domains, i.e C_2 is preceeded by an aminotransferase (AMT_1) domain (Table 4-4), and thus suggested to be a hybrid C domain (Rausch et al., 2007) catalyzing the condensation reaction between an upstream aminated polyketide and a downstream amino acyl-S-PCP_2. The C_4, C_5 and C_8 domains are preceeded by E_3, E_4 and E_7 domains, respectively and are expected to belong to the class of ^DC_L domains (Table 4-4). These ^DC_L domains catalyse the condensation reaction between an upstream D-amino acyl or petidyl-S-PCP with a downstream L-amino acyl-S-PCP. The ^DC_L domains i.e C_4, C_5 and C_8 differs concerning some critical residues from those of the ^LC_L domains. This allows to differentiate their respective function (Rausch et al., 2007) (Figure 4-11).
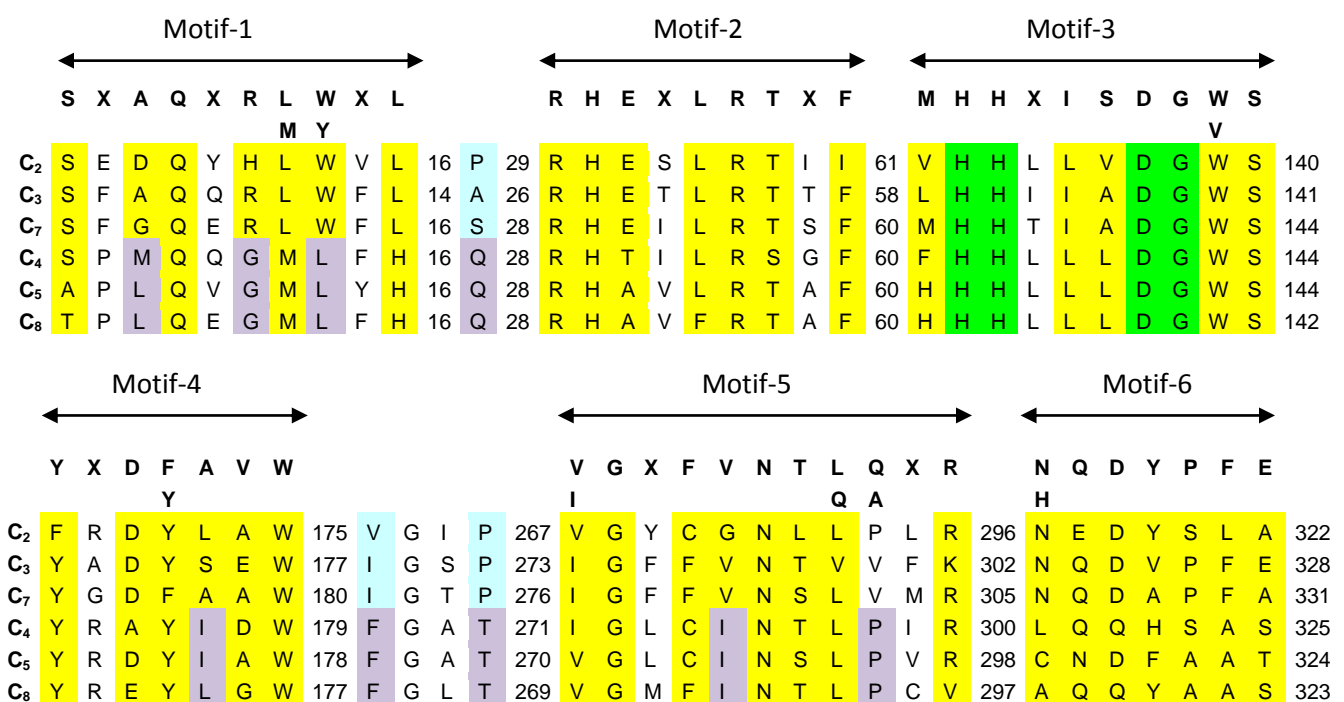
Figure 4-11: Characteristic motifs in the C domains of PKS/NRPS gene locus-1 are highlighted in yellow. The catalytically active residues are highlighted in green. The residues critical for the ^DC_L domains are highlighted in violet and those of ^LC_L domains are highlighted in blue (Rausch et al., 2007). The conserved residues from the literature (Konz and

Marahiel, 1999) are highlighted in bold. The numbers indicate the position of amino acids. Suffices indicate the respective modules.

Ketosynthase domains

After loading the substrates to the respective ACP domains of PKS modules, the KS domain links all the substrates sequentially starting from the intitation module by Claisen condensation (Staunton and weissman, 2001; Hertweck, 2009). Since the two PKS modules, with KS_1 and KS_6 are elongating modules, both of them require KS domains for this activity. PKS module-6 (Haur_1860) is preceeded by an NRPS module (Haur_1859) (Table 4-4). Therefore, KS_6 is likely to mediate Claisen condensation between an upstream peptidyl-S-PCP_5 and a downstream malonyl-S-ACP_6. Sequence alignment of both the KS domains, i.e KS_1 and KS_6 showed that the catalytic triad 'C/H/H'(Zhang et al., 2006) is conserved in KS_1 and KS_6 (Figure 4-12). The histidine residues are necessary for malonyl-CoA decarboxylation and the cysteine moiety is required for the condensation reaction between the resulting carbanion and the upstream acyl-S-ACP or peptidyl-S-PCP (Carlson et al., 2010; Pawlik et al.,2007).

```
            C              H              H
KS₁    C  S  S  173  H  G  T  306  G  H  L  343
KS₆    C  S  T  175  H  G  T  311  G  H  L  348
```

Figure 4-12: Catalytic cysteine/histidine/histidine motif (highlighted in green) extracted from the KS domains of PKS/NRPS gene locus-1. Consensus sequences from the literature (Zhang et al., 2006) are highlighted in bold. The numbers indicate the position of the amino acid. Suffices indicate the respective modules.

Epimerization domains

One of the unusual features of non-ribosomal peptides is the occasional presence of D-amino acids as part of the peptidic structure (Schwarzer et al., 2003). In NRPS megasynthetases, the epimerization reaction is performed by a so called E domain. In rare cases the C domain might also perform the epimerization reaction, often referred to as C/E dual domain (Vallet-Gely et al., 2009). Sequence alignments of the epimerization domains E_3, E_4 and E_7 from the hybrid gene locus-1 (Table 4-4) showed that all seven core motifs (Konz and Marahiel, 1999) were conserved with only a slight variation (Figure 4-13).
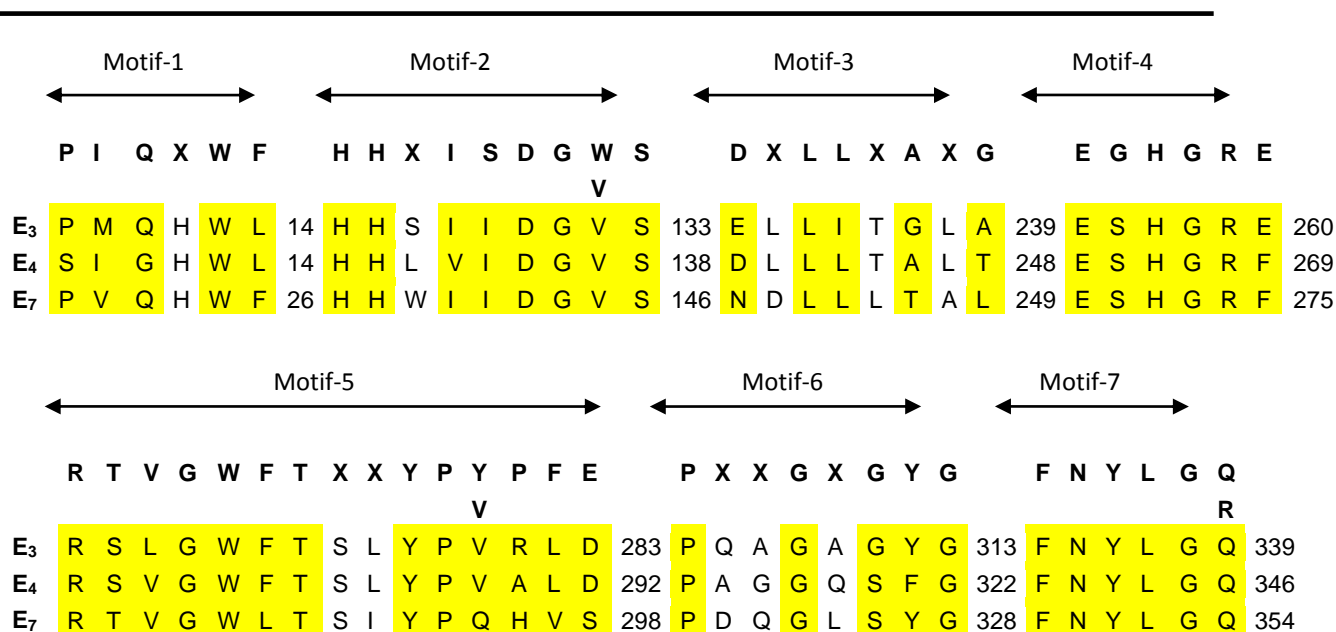
Figure 4-13: Characteristic motifs in E domains (highlighted in yellow) of PKS/NRPS gene locus-1. Consensus sequences from the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers indicate the position of amino acids. Suffices indicate the respective modules.

Methyltransferase domain

N-methylation and C-methylation is one of the most common modification reactions in NRPS and PKS systems, and occurs at the thioester bound stage (Schwarzer et al., 2003, Hertweck, 2009). It is the methyltransferase (MT), which transfers the methyl group from S-adenosyl-methionine to the acceptor. All MTs are expected to possess the typical glycine rich S-adenosyl-methionine binding motif-1, (V/I/L)-(L/V)-(D/E)-(V/I)-G-(G/C)-G-(T/P)-G (Konz and Marahiel, 1999). PKS/NRPS gene locus-1 is encoding for one MT, i.e MT_2 (Table 4-4). BLASTP of the MT_2 domain protein sequence showed 41% identity to Nda A from *Nodularia spumigena.* The MT_2 domain amino acid sequence also contains additional five core motifs typical for N-MT (Konz and Marahiel, 1999; Hacker et al., 2000). Except for motif-5 and motif-3 all other motifs are conserved, in MT_2 with respect to reference sequences (Figure 4-14).

Figure 4-14: Conserved motifs for the N-MT (highlighted in yellow) domain MT_2 of PKS/NRPS gene locus-1, NdaA from *Nodularia spumigena* (acc.no-AA064403.1), MycA from *Anabaena sp.* 1tu44S16 (acc no-ABW96246.1), Pro-D from *Planktothrix rubescens* (acc no-ACG63858.1). The conserved motifs as described by Hacker et al., 2000 are referred to in roman numbers, whereas the motifs taken from Konz and Marahiel, 1999 are referred to in numeric. The numbers indicate the position of amino acids. Suffices indicate the respective modules.

Ketoreductase domain

Ketoreductases are NADPH dependant enzymes catalyzing the reduction of the β-keto-group of the keto acyl-*S*-ACP to a hydroxyl group (Hertweck, 2009). The KR_6 domain is the only KR domain in the PKS/NRPS gene locus-1 (Table 4-4) and carries the NADP(H) binding site G-X-G-X-X-G(A)-X-X-X-A which is necessary for its activity (Bachmann and Ravel, 2009). The KR_6 domain also includes the catalytic triad S-136/Y-149/N-153 required for the catalytic activity (Pawlik et al., 2007) (Figure 4-15). Sequences of KR domains can also reveal the configuration of the chiral centre formed by the reduction reaction. However, the KR_6 domain amino acid sequence lacks in both, the consensus motifs LDD, P-144 and N-148 required for the R-

configuration with respect to the hydroxyl group (B type) and W-141 needed for the S-configuration with respect to the hydroxyl group (A type) (Caffrey., 2003).

$$\text{G} \quad \text{X} \quad \text{G} \quad \text{X} \quad \text{X} \quad \text{G} \quad \text{X} \quad \text{X} \quad \text{X} \quad \text{A} \qquad \text{S} \qquad \text{Y} \qquad \text{N}$$

A

KR$_6$ G N G G L G R Q I A 16 S 136 Y 149 N 153
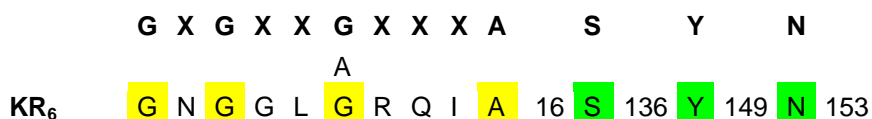
Figure 4-15: Characteristic motifs in the KR$_6$ domain (highlighted in yellow) of PKS/NRPS gene locus-1. The catalytic active residues are highlighted in green. Consensus sequences from the literature (Bachmann and Ravel, 2009; Pawlik et al., 2007) are highlighted in bold. The numbers indicate the position of amino acid. The suffix indicates the respective module.

Aminotransferase domain

The PKS/NRPS hybrid module-1 includes a putative AMT$_1$ domain, the encoding sequences for which are located between the PKS and NRPS part of modules- 1 and 2 respectively in Haur_1857 (Table 4-4). Aminotransferases catalyse the transfer of an amino group from an amino acid to an α-keto acid in the presence of pyridoxal phosphate. Therefore, an AMT is expected to posses the conserved residues K-277 for the binding of the pyridoxal phosphate cofactor, and both, E-220 and D-248 for interacting with the 3-OH and the N1 of pyriodixal phosphate (Naranjo et al., 2005) (Figure 4-16). Alignment of the AMT$_1$ domain protein sequences with mycosubtilin synthetase A showed the presence of pyridoxal phosphate binding site sequences (Duitman et al., 1999) (Figure 4-16). Therefore, the AMT$_1$ domain is suggested to aminate the ketoacyl-S-ACP$_1$.

AMT$_1$	E	333	D	366	D	L	A	T	Y	G	K	I	V	G	G	G	M	P	I	G	V	V	A	G	R	407
MycA	E	354	D	387	D	L	V	T	Y	G	K	V	I	G	G	G	Q	P	L	G	V	V	A	G	K	428
BmyD	E	347	D	380	D	L	V	T	Y	G	K	I	I	G	G	G	Q	P	L	G	I	V	A	G	K	421
iturin a syn a	E	375	D	408	D	L	V	T	Y	G	K	V	I	G	G	G	L	P	I	G	I	V	G	G	K	449
MicA	E	382	D	415	D	I	A	T	Y	G	K	V	I	A	G	G	M	P	I	G	V	V	A	G	K	456

Figure 4-16: Conserved pyridoxal phosphate binding site sequences (Duitman et al., 1999) of AMT$_1$ in PKS/NRPS gene locus-1, with functional AMT domains are highlighted in yellow. The residues typical for the binding of pyridoxal phosphate are highlighted in orange (Naranjo et al., 2005). The numbers indicate the position of amino acids. The suffix indicates the respective module. MycA from *Desmospora sp.* 8437 (Acc no-ZP08466280), BmyD from *Bacillus subtilis* (Acc no-AAN07012.1), iturin a synthetase a from *Paenibacillus elgii* B69 (Acc no-ZP09075091.1), MicA from *Planktothrix rubescens* NIVA-CA-98 (Acc no- CA048259.1).

Thioesterase domain

Thioesterases belong to the family of α/β-hydrolases catalyzing the cleavage of thioesters. TE domains must have the conserved motif 1, i.e G-X-S-X-G, which is typical for hydrolytic enzymes (Rajakumari and Daum, 2009). In addition, TE amino acid sequences comprise a highly conserved aspartate residue, which is 25 residues downstream to the conserved motif 1. Another conserved region is motif-2, i.e G-X-H (Konz and Marahiel, 1999). The latter motif is at approximately 130 residues downstream to the aspartate residue. The catalytic triad S/D/H extracted from the above consensus motifs is necessary for the TE domain activity (Bruner et al., 2002). The C-terminus of the PKS/NRPS gene locus-1 (Table 4-4) encodes a TE domain as a part of *haur_1861*. Examining the amino acid sequences of this TE domain revealed the presence of the highly conserved catalytic triad with S, D and H motifs (Figure 4-17). Therefore, this TE domain was suggested to be active in releasing the fully assembled hybrid product from the megasynthetase. Moreover, the TE domain includes a characteristic proline residue 64 residues upstream to motif-1, which is typical for TE acting as cyclases (Roonsawang et al., 2007) (Figure 4-17). So, this TE domain of PKS/NRPS gene locus-1 is speculated to cyclize the linear putative lipopeptide by mediating a nucleophilic attack between the amino group introduced by the action of AMT_1 on the fatty acid and the terminal carboxyl group.

			G	**X**	**S**	**X**	**G**		**D**		**G**	**X**	**H**	
TE	P	7	G	H	S	F	G	75	D	100	G	D	H	237
CrpD	P	7	G	H	S	F	G	75	D	100	G	N	H	244
BA TE	P	7	G	H	S	F	G	75	D	100	G	D	H	248
MF TE	P	7	G	H	S	F	G	75	N	100	G	D	H	244
CC TE	P	7	G	H	S	F	G	75	D	100	G	D	H	240

Figure 4-17: Consensus sequences of the TE domain (highlighted in yellow) from the PKS/NRPS gene locus-1; CrpD from *Nostoc* sp. ATCC 53789 (Acc no-ABM21572.1); BA, TE from *Burkholderia ambifera* IOP40-10 (Acc no- ZP02892872.1); MF TE, *Myxococcus fulvus* HW-1 (Acc no- YP04668982.1); CC TE, *Cyanotheca* sp. CCY0110 (Acc no- ZP01728834.1). The catalytic active site residues are highlighted in green. The proline typical for the cyclase group of TE is highlighted in blue. Conserved sequence from the literature (Konz and Marahiel, 1999) is highlighted in bold. The numbers indicate the position of amino acids.

Putative structure encoded by hybrid gene locus-1

Based on PKS/NRPS gene locus-1 sequence, the locus is predicted to encode a putative lipopeptide composed of β-amino fatty acid tail linked to a hexapeptide (Figure 4-18). The latter should consist of six amino acids including D-asparagine, D-threonine, L-serine, D-ornithine, L-glycine and an unpredictable amino acid (Figure 4-8B). Ornithine cyclodeaminase encoded by *haur_1863* (Table 4-4) is expected to convert ornithine loaded on PCP_7 to proline. Finally, the molecule is supposedly cyclised by an intramolecular attack of the β-amino group of the fatty acid side chain with the terminal carboxy group of L-glycine to form a lactam ring (Figure 4-18). Similar to the other lipopeptides like mycosubtilin, bacillomycin and iturin, the putative lipopeptide encoded by the hybrid gene locus-1 is also expected to possess antimicrobial activity (Duitman et al., 1999).

(A)

(B)
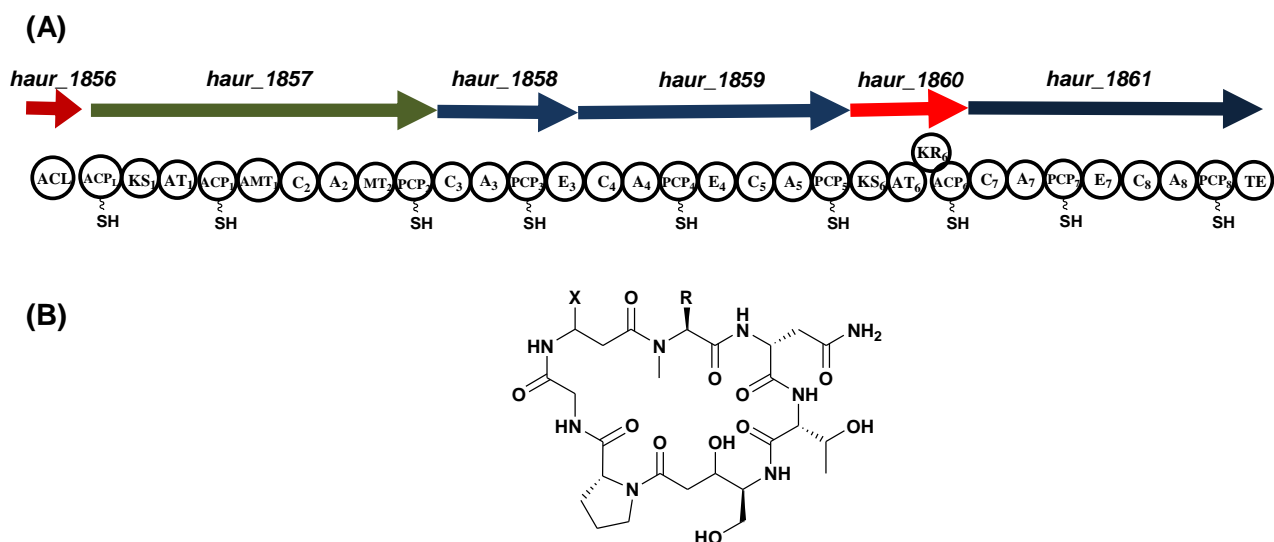
Figure 4-18: PKS/NRPS gene locus-1

(A) Putative PKS (red) and NRPS (blue) genes in PKS/NRPS gene locus-1. Brown, acyl-CoA synthetase; green, PKS/NRPS hybrid.

(B) Predicted strcture of the putative lipopeptide encoded by PKS/NRPS gene locus-1. X, Unknown fatty acid side chain; R, Unknown amino acid.

4.1.4.1.2 PKS/NRPS gene locus-2

The PKS/NRPS gene locus-2 with eleven NRPS and a single PKS module spanning from haur_1871- haur_1888 is 65.3 Kb in size (Table 4-5). The biosynthesis of the putative compound is most likely to start with the activation and loading of an unusual long chain starter unit by the action of AMP-dependent synthetase and ligase (Figure

4-19) (Haur_1884) to the ACP_L encoded by *haur_1873.* The loaded starter is then proposed to condense with the malonyl- S-ACP_1 linked to the partially reducing PKS module-1 in Haur_1873. In addition to Smith and Tsai (2007) model, the selection of malonyl-CoA by the AT_1 domain is supported by the presence of the consensus motifs GQSVG and IAFH (Yadav et al., 2003) in the AT_1 domain amino acid sequence (Figure 4-20). The PKS module-1 includes an internal MT_1 domain (Table 4-5). BLASTP of the MT_1 encoded by *haur_1873* showed 52% identity to *KtzH* (*O*-MT) of *Kutzneria sp.* 744 (Fujimori et al., 2007). Therefore, the *O*-MT_1 of *haur_1873* was presumed to methylate the hydroxyl group generated collectively by the action of KR_1 and DH_1 domains on the starter unit.

Gene locus	Modules	Putative functions
haur_1871		p-hydroxymandelate synthase
haur_1872		Hypothetical protein
haur_1873	Module-L Module-1	ACP_L KS_1-AT_1-DH_1-MT_1-KR_1-ACP_1
haur_1874	Module-2	C_2-A_2-PCP_2-E_2
haur_1875	Module-3	C_3-A_3-PCP_3-E_3-$C_{4'}$
haur_1876		Taurine catabolism dioxygenase (TfdA/TauD)
haur_1877	Module-4	C_4-A_4-PCP_4-E_4
haur_1878		$C_{5'}$
haur_1879	Module-5 Module-6	C_5-A_5-PCP_5 C_6-A_6-PCP_6
haur_1880		Hypothetical
haur_1881	Module-7	C_7-A_7-PCP_7-$C_{8'}$
haur_1882	Module-8 Module-9 Module-10 Module-11 Modue-12	C_8-A_8-PCP_8 C_9-A_9-PCP_9 C_{10}-A_{10}-PCP_{10} C_{11}-A_{11}-PCP_{11} C_{12}-A_{12}-PCP_{12}-TE_1
haur_1883		Phenylalanine/histidine ammonia-lyase
haur_1884		AMP-dependant synthetase and ligase (AMP)
haur_1885		TE_2

haur_1886		Glycosyl transferase
haur_1887		L-p-hydroxyphenylglycine (HPG) transaminase
haur_1888		4-hydroxyphenylpyruvate dioxygenase

Table 4-5: List of genes in PKS/NRPS gene locus-2 and their putative functions. Suffices indicate the respective modules. L, Loading module; ACP, Acyl Carrier Protein; KS, Ketosynthase; AT, Acyltransferase; DH, Dehydratase; MT, Methyltransferase; C, Condesation; A, Adenylation; PCP, Peptidyl Carrier Protein; E, Epimerization. The genes encoding the biosynthesis of HPG are highlighted in blue.

Figure 4-19: Characteristic motifs of acyl-CoA synthetase in the AMP dependent synthetase and ligase encoding genes (highlighted in green) in PKS/NRPS gene loci-2, 3, 4 and 5 and in the NRPS gene locus-4 of *H. aurantiacus* DSM 785. The histidine residue is typical for long chain acyl-CoA synthetases and is highlighted in green. The numbers indicate the position of amino acids. Conserved sequences from the literature (Gulick, 2007) are highlighted in bold.

(A) AT domains

					Q				G	H	S	*	G			R				#	H			§			$			V	
Hybrid-2-AT₁	Q	11	Q	52	G	Q	S	V	G		82	R		105	I	A	F	H	194	T	224	N	243	F	248
Hybrid-3-AT₁	Q	11	Q	63	G	H	S	V	G		93	R		116	M	A	Y	H	195	T	225	Q	244	V	249
Hybrid-4-AT₁	G	11	Q	52	G	H	S	L	G		82	R		105	I	A	F	H	194	T	224	N	243	F	248
Hybrid-4-AT₅	Q	14	L	65	G	H	S	L	G		95	R		118	V	A	A	H	192	N	222	H	242	V	247
PKS-AT	-		Q	44	G	H	S	L	G		74	R		97	H	A	F	H	175	T	205	Q	223	V	228

(Motif headers for AT domains: Q, Q, G H S * G, R, # H, §, $, V)

(B) KS domains

	C				H				G	H		
Hybrid-2-KS₁	C	S	G	173	H	G	T	308	G	H	T	345
Hybrid-3-KS₁	C	S	S	168	H	G	T	303	G	H	L	338
Hybrid-4-KS₁	C	S	G	172	H	G	T	307	G	H	S	346
Hybrid-4-KS₅	C	S	T	172	H	G	T	308	G	H	L	345
PKS-KS	C	S	S	209	H	G	T	342	G	H	T	380

(C) ACP domains

	G	X	X	S		
Hybrid-2-ACP_L	G	L	S	S	I	32
Hybrid-2-ACP₁	G	L	D	S	I	32
Hybrid-3-ACP_L	G	A	T	S	L	32
Hybrid-3-ACP₁	G	L	D	S	L	32
Hybrid-4-ACP_L	Q	I	D	S	I	32
Hybrid-4-ACP₁	G	L	D	S	I	32
Hybrid-4-ACP₅	G	G	D	S	L	30

(D) KR domains

Motif header: G X G X X G X X X A (A), S, Y, N

	G	X	G	X	X	G	X	X	X	A		S		Y		N	
Hybrid-2-KR₁	G	L	S	G	I	G	L	E	V	A	17	S	151	Y	165	N	169
Hybrid-3-KR₁	G	T	G	G	V	G	A	Q	L	A	15	S	134	Y	150	N	154
Hybrid-4-KR₁	G	S	G	G	L	G	L	E	I	A	17	S	145	Y	159	N	163
Hybrid-4-KR₃	G	L	G	G	I	G	S	Q	F	A	17	S	159	Y	173	N	177
PKS-KR	G	G	K	G	I	T	A	E	S	V	22	S	141	Y	155	N	159

(E) DH domains

	L	X	X	H	X	X	X	G	X	X	X	X	P	
Hybrid-2-DH₁	L	R	D	H	Q	V	Q	G	V	V	V	L	P	46
Hybrid-4-DH₁	L	A	D	H	V	V	Q	E	Q	V	L	L	P	46

Figure 4-20: Characteristic motifs in AT, KS, ACP, KR and DH domains of PKS modules extracted from PKS/NRPS gene loci- 2, 3 & 4 and the PKS gene locus-1 in *H. aurantiacus* DSM 785 (highlighted in yellow). Catalytically active site residues are highlighted in green. Consensus sequences from the literature (Yadav et al., 2003; Pawlki et al., 2007; Bachmann and Ravel, 2009) are highlighted in bold. * indicates the invariant residues Q/M/I/L/V/F/A/M, # indicates S/F/P, § indicates T/A/N/G/E/D/S and $ indicates N/S/H/Q. The numbers represent the position of amino acids. Suffices indicate the respective modules.

Amino acid sequences of the eleven NRPS A domains (Table 4-5), i.e A_2-A_{12} from PKS/NRPS gene locus-2 include ten conserved core motifs (Figure 4-22A). The Stachelhaus code extracted from these core motifs was used to predict the substrate of the respective A domains. Except for A_6, A_{10} and A_{12}, the Stachelhaus code of all other A domains were at least 70% identical with of those known substrates (Figure 4-22B). The latter are highly likely to activate the predicted substrates.

The Stachelhaus code extracted from the A_2 domain of the PKS/NRPS gene locus-2 suggested that it activates the non-proteinogenic amino acid 4-hydroxyphenylglycine (HPG). The PKS/NRPS gene locus-2 includes the entire set of genes, *haur_1871, haur_1887* and *haur_1888* (Table 4-5) encoding p-hydroxymandelate synthase, p-hydroxyphenylglycine transaminase and 4-hydroxyphenylpyruvate dioxygenase, respectively, required for the synthesis of HPG (Figure 4-21) (Kastner et al., 2012). This unusual amino acid is also one of the building blocks of bioactive compounds like vancomycin, in which this amino acid plays an important role for its biological function (Hubbard et al., 2000).
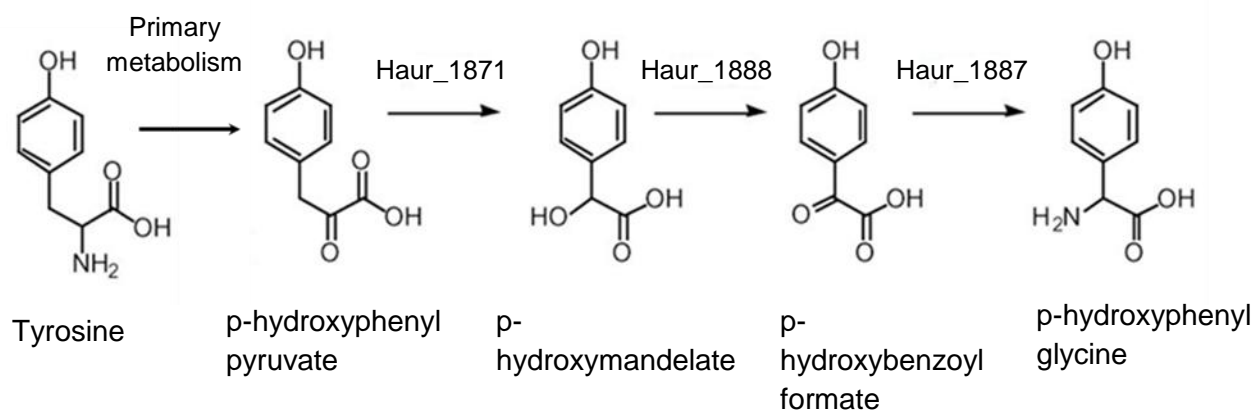
Figure 4-21: Biosynthesis of p-hydroxyphenylglycine (HPG).

Results

Motif-1 — Motif-5

Consensus:
- Motif-1: L S Y X E L
- Motif-2: L K A G X A Y L P I D (V)
- Motif-3: L A Y X X Y T S G T T G X P K G (S)
- Motif-4: F D X S
- Motif-5: N X Y G P T E

	Motif-1		Motif-2		Motif-3		Motif-4		Motif-5	
A2	Y S Y A E L	52	F K A G G L F L P L D	102	L A Y M I Y T S G T T G T P K A	181	F D I F	218	V L Y G P T E	318
A3	L S Y G E L	51	L K A G A A Y V P L D	101	L V Y L I Y T S G S T G R P K A	181	F D A S	218	N A Y G P T E	308
A4	L S Y A E L	51	L K A G G A Y V P L D	101	L M Y I I Y T S G S T G Q P K G	181	A D L G	218	N H Y G P T E	315
A5	L T Y Q E L	52	L K A G A A Y L P I D	102	L A Y I I Y T S G S T G K P K G	181	A D L G	218	N H Y G P T E	315
A6	L S Y A E L	51	L K A G A G Y V P L D	101	P V Y T I Y T S G S T G T P K G	181	F D F G	218	N G Y G P T E	314
A7	L S Y H A L	51	L K A G A A Y A P V D	101	L A Y V M Y T S G S T G R P K G	181	F D A S	217	N G Y G P T E	312
A8	L S Y H A L	51	L K A G A A Y A P V D	101	L A Y V M Y T S G S T G R P K G	181	F D A S	217	N G Y G P T E	315
A9	L S Y A E L	45	L K A G G A Y V P L D	95	L A Y M I Y T S G S T G K P K G	181	F D V S	218	N L Y G P T E	315
A10	W T Y A Q M	50	L K A G G A Y L P L D	100	L A Y V I Y T S G S T G V P K G	181	F D I S	218	N M Y G P T E	311
A11	L S Y G A L	45	M K A G G A Y V P L D	95	L A Y M I Y T S G S T G T P K G	181	F D V S	218	N L Y G P T E	315
A12	W S Y A E L	50	I K A G A A Y L P L D	100	L A Y V I Y T S G S T G V P K G	181	F D I S	218	N M Y G P T E	311

Motif-6 — Motif-10

Consensus:
- Motif-6: G E L X I X G X G L A R G Y L (K)
- Motif-7: Y R T G D L
- Motif-8: G R X D X Q V K I R G X R I E L G E I E
- Motif-9: L P X Y M V P (I)
- Motif-10: N G K V D R (L)

	Motif-6		Motif-7		Motif-8		Motif-9		Motif-10	
A2	G E L Y I G G A G V S R G Y L	378	Y R T G D L	403	G R I D Q Q V K I R G F R I E L G E I E	436	L P E Y M L P	490	N G K V D R	510
A3	G E L Y I A G P G L A W G Y L	377	Y R S G D L	396	G R V D H Q V K I R G F R I E T G E I E	429	L P E H M L P	483	N G K L D R	503
A4	G E L Y V G G A A L A R G Y W	378	Y R T G D V	408	G R A D D Q V K I R G F R I E L G E I E	441	L P D Y M V P	495	N G K L D R	515
A5	G E L Y I G G A G V A R G Y L	378	Y R S G D V	408	G R S D D Q V K I R G Y R V E L G E I S	441	Q Q G I F A E	497	N R H M T V	515
A6	G E L C V A G P S L S Q G Y L	387	Y R T G D V	406	G R I D H Q V K V R G Y R I E L G E I E	439	L P G Y M L P	489	N G K I N R	509
A7	G E L Y L G G A G L A R G Y L	383	Y R S G D L	403	G R R D Q Q V K V R G Y R I E L E E I V	436	V P S Y M L P	487	N G K V D R	507
A8	G E L Y L G G A G L A R G Y L	383	Y R S G D L	403	G R R D Q Q V K V R G Y R I E L E E I V	436	V P S Y M L P	487	N G K V D R	507
A9	G E L Y I G G L Q V G R G Y Y	375	Y R T G D L	405	G R I D H Q V K V R G L R I E L G E I E	438	L P E F M L P	491	N G K L D R	511
A10	G E L Y I G G S G I A R G Y L	367	Y R T G D L	397	G R I D H Q V K V R G F R I E L G E I E	430	L P E Y M L P	483	N G K V D R	503
A11	G E V Y I G G V Q V G R G Y H	375	Y R T G D V	405	G R I D H Q V K L R G L R I E L G E I E	438	L P E Y M L P	491	N G K V D R	511
A12	G E L Y I G G S G I A R G Y L	367	Y R T G D L	397	G R I D H Q V K V R G F R I E L G E I E	430	L P E Y M L P	483	N G K V D R	503

(B)

Gene locus	A domain	Stachelhaus code	Predicted substrates
haur_1874	A_2	DIFHFGLIIK	HPG
haur_1875	A_3	DAFTIAAVWK	Phenylalanine
haur_1877	A_4	DLTKIGHVGK	Asparagine
haur_1879	A_5	DLTKVGHVG*	Aspartate
haur_1879	A_6	DFFHHGGVNK	Unknown
haur_1881	A_7	DAFWLGGTFK	Valine
haur_1882	A_8	DAFWLGGTFK	Valine
haur_1882	A_9	DVWHISLVDK	Serine
haur_1882	A_{10}	DITQIGMVYK	Unknown
haur_1882	A_{11}	DVWHLSLVDK	Serine
haur_1882	A_{12}	DIVQIGMVYK	Unknown

Figure 4-22: PKS/NRPS gene locus-2 A domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS A domains (highlighted in yellow) from PKS/NRPS gene locus-2. The 10 core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules.

(B) The table lists the Stachelhaus codes extracted from the sequences of PKS/NRPS gene locus-2 A domains and their putative substrates. HPG, 4-hydroxyphenylglycine. The asterisk indicates the unknown amino acid.

The NRPS modules of the PKS/NRPS gene locus-2 contain supposedly superfluous C domains (Table 4-5). Aligning the amino acid sequences of all C domains (Figure 4-23A), i.e C_2-C_{12}, $C_{4'}$, $C_{5'}$ and $C_{8'}$ from the hybrid gene locus-2 revealed that C_4, C_5 and $C_{8'}$ did not posses the consensus histidine motif in their active sites (Bergendahl et al., 2002; Roche et al., 2008). Therefore, these C domains are presumed to be inactive. However, these inactive C domains were either succeeded or preceeded by an active C domain, i.e $C_{4'}$, $C_{5'}$ and C_8 (Table 4-5). Thus, all the NRPS modules of the PKS/NRPS gene locus-2 are proposed to be active in the biosynthesis (Figure 4-19, Figure 4-23, Figure 4-24).

The E_2, E_3 and E_4 domains of the PKS/NRPS gene locus-2, include the seven conserved motifs (Figure 4-23B) (Konz and Marahiel, 1999). Therefore, these E domains were presumed to convert the L-HPG-*S*-PCP_2, L-Phe-*S*-PCP_3 and L-Asn-*S*-PCP_4 into D-HPG-*S*-PCP_2, D-Phe-*S*-PCP_3 and D-Asn-*S*-PCP_4.

(A) C domains

Motif-1

	S	X	A	Q	X	R	L	W	X	L		R	H	E	X	L	R	T	X	F		M	H	H	X	I	S	D	G	W	S	
							M		Y																					V		
C₂	C	F	A	Q	E	R	L	W	L	M	16	R	H	E	I	L	R	T	T	F	61	L	H	H	I	V	S	D	D	W	S	144
C₃	S	P	L	Q	Q	G	L	L	F	H	16	R	H	A	L	R	T	V	F		61	N	H	H	L	L	L	D	G	W	S	145
C₄'	T	P	T	Q	Q	G	M	L	F	H	16	Q	H	T	I	L	R	S	C	F	61	Q	H	H	I	L	L	D	G	W	C	145
C₄	S	P	Q	Q	Q	H	V	W	A	L	21	R	Y	E	I	L	R	T	T	F	65	V	P	T	L	C	A	D	N	L	S	143
C₅	S	P	Q	Q	K	Q	L	W	S	L	21	R	Y	E	I	L	R	T	Q	F	65	L	P	A	L	T	A	D	L	T	T	141
C₅'	S	P	M	Q	K	G	M	L	F	H	16	R	H	S	V	L	R	T	A	F	61	R	H	H	L	L	L	D	G	W	S	145
C₆	S	F	T	Q	E	R	Y	W	I	L	16	R	H	A	I	L	R	T	T	F	61	M	H	H	I	I	T	D	D	W	S	144
C₇	S	F	A	Q	E	R	L	W	F	L	16	R	H	A	S	F	R	T	T	F	62	C	H	H	I	I	S	D	G	W	S	148
C₈'	T	S	A	Q	Q	R	V	W	F	F	16	R	H	E	I	L	R	A	S	F	61	I	Q	H	S	I	T	D	A	W	S	144
C₈	S	F	A	Q	E	R	L	W	F	L	16	R	H	A	S	F	R	T	T	F	62	C	H	H	I	I	S	D	G	W	S	148
C₉	S	F	A	Q	E	R	L	W	F	L	16	R	H	S	L	R	T	T	F		62	L	H	H	S	I	T	D	G	W	S	141
C₁₀	S	F	A	Q	E	R	L	W	F	L	16	R	H	A	S	L	R	T	S	F	62	L	H	H	T	I	T	D	G	W	S	141
C₁₁	S	F	A	Q	E	R	L	W	F	L	16	R	H	A	S	L	R	T	T	F	62	L	H	H	T	I	T	D	G	Q	S	141
C₁₂	S	F	A	Q	E	R	L	W	F	L	16	R	H	A	S	L	R	T	T	F	62	L	H	H	T	I	T	D	G	Q	S	141

Motif-4	Motif-5	Motif-6

	Y	X	D	F	A	V	W		V	G	X	F	V	N	T	L	Q	X	R		N	Q	D	Y	P	F	E	
				Y							I						Q	A				H						
C₂	V	W	Q	-	R	Q	R	184	I	G	C	F	I	N	M	L	V	L	R	298	N	A	E	V	P	F	E	330
C₃	Y	R	D	Y	I	A	W	180	V	G	L	F	I	N	T	I	P	V	C	298	Y	H	Y	T	G	L	N	330
C₄'	Y	R	E	Y	I	A	W	156	L	G	L	F	I	N	T	L	P	V	R	298	Y	E	Y	T	P	V	A	330
C₄	E	W	Q	-	H	E	L	178	L	G	L	F	A	K	A	V	P	L	A	292	W	Q	E	S	F	H	W	331
C₅	E	W	L	-	N	D	T	162	L	G	L	F	A	K	H	V	P	F	F	279	W	Q	E	Y	W	A	W	328
C₅'	Y	G	E	Y	I	A	W	186	V	G	L	F	I	N	T	L	P	L	R	301	Y	E	Y	S	S	L	V	333
C₆	H	W	Q	-	R	Q	W	185	I	G	F	F	L	N	T	L	P	L	R	299	H	Q	D	L	P	F	E	331
C₇	V	W	Q	-	R	Q	R	189	I	G	F	L	V	N	T	W	T	L	R	302	H	R	D	L	P	F	E	334
C₈'	A	W	Q	-	A	A	W	186	L	G	F	F	I	N	T	L	V	L	R	299	Q	Q	D	L	P	L	E	331
C₈	V	W	Q	-	R	Q	R	189	I	G	F	L	V	N	T	W	T	L	R	302	H	R	D	L	P	F	E	334
C₉	L	W	Q	-	R	A	V	182	I	G	F	F	A	N	T	L	V	V	R	295	Q	Q	E	L	P	F	E	327
C₁₀	V	W	Q	-	R	A	V	182	I	G	F	F	A	N	T	L	V	V	R	295	H	Q	E	L	P	F	E	327
C₁₁	V	W	Q	-	R	A	V	182	I	G	F	F	A	N	T	L	V	V	R	295	H	Q	E	L	P	F	E	327
C₁₂	V	W	Q	-	R	A	V	182	I	G	F	F	A	N	T	L	V	V	R	295	H	Q	E	L	P	F	E	327

(B) E domains

Motif-1	Motif-2	Motif-3	Motif-4

	P	I	Q	X	W	F		H	H	X	I	S	D	G	W	S		D	X	L	L	X	A	X	G		E	G	H	G	R	E	
															V																		
E₅	P	I	Q	Q	W	F	14	H	H	L	V	F	D	G	I	S	142	E	V	L	L	T	A	L	A	256	E	S	H	G	R	A	285
E₆	P	I	Q	Q	W	F	14	H	H	L	I	F	D	G	V	S	143	D	I	L	L	T	A	L	A	259	E	S	H	G	R	E	282
E₇	P	I	Q	H	W	F	14	H	H	L	A	I	D	T	V	S	142	D	A	L	L	A	A	L	T	259	E	G	H	G	R	E	282

	Motif-5																Motif-6								Motif-7						
	R	T	V	G	W	F	T	X	X	Y	P	Y	P	F	E		P	X	X	G	X	G	Y	G		F	N	Y	L	G	Q	
												V																			R	
E₅	R	T	I	G	W	F	T	S	L	Y	P	V	I	L	D	307	P	E	Y	G	L	S	Y	G	337	F	N	Y	L	G	Q	356
E₆	R	T	I	G	W	F	T	A	I	A	P	L	R	L	T	305	P	Q	H	G	V	G	Y	G	335	F	N	Y	L	G	Q	357
E₇	R	T	V	G	W	F	T	S	R	F	P	V	L	L	Q	305	P	Q	R	G	I	G	Y	G	335	F	N	Y	L	G	T	364

(C) PCP domains

	D	X	F	F	X	L	G	G	H	S	L	
									D		I	
PCP₂	D	N	F	F	E	L	G	G	D	S	I	30
PCP₃	D	N	F	F	E	L	G	G	D	S	I	30
PCP₄	N	N	F	F	E	L	G	G	D	S	I	30
PCP₅	D	K	F	F	D	L	G	G	H	S	L	30
PCP₆	D	N	F	F	H	I	G	G	H	S	L	30
PCP₇	T	N	F	F	Q	V	G	G	H	S	L	30
PCP₈	T	N	F	F	Q	V	G	G	H	S	L	30
PCP₉	S	N	F	F	Q	L	G	G	H	S	L	30
PCP₁₀	A	H	F	F	Q	L	G	G	H	S	L	30
PCP₁₁	A	H	F	F	Q	L	G	G	H	S	L	30
PCP₁₂	Q	N	F	F	E	L	G	G	H	S	M	30

Figure 4-23: Characteristic motifs in C (A), E (B) and PCP (C) domains (highlighted in yellow) from the NRPS modules of PKS/NRPS gene locus-2. The catalytically active site residues are highlighted in green. Consensus sequences from the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers indicate the position of amino acids. Suffices indicate the respective modules. The inactive domains are highlighted in grey.

The locus harbors two TE domains (Table 4-5) i.e TE₁ and TE₂. The TE₁ encoded on *haur_1874* contains the catalytic triad S-89/D-116/H-264 required for activity. In addition, TE₁ includes a glycine residue at position 27, typical for the hydrolase type of TEs (Figure 4-24) (Roongsawang et al., 2007). The TE₂ of uncertain function on *haur_1885* contains S-89/S-116/H-264. As per the surfactin synthetase model, the D-116 of the TE domain was required for the activity (Bruner et al., 2002). But in TE₂ domain amino acid sequence, serine instead of aspartate (Figure 4-24) raises question about the activity and role of this TE domain in the biosynthesis. Interestingly the TE₂ domain amino acid sequence from the PKS/NRPS gene loci-3, 4 and 5 of *H. auaratiacus* DSM 785 also harbors the serine amino acid instead of aspartate (Figure 4-24).

			G	**X**	**S**	**X**	**G**		**D**		**G**	**X**	**H**	
Hybrid-2-TE₁	G	11	G	W	S	F	G	75	D	100	G	D	H	232
Hybrid-4-TE₁	G	11	G	W	S	M	G	75	D	100	G	T	H	235
Hybrid-5-TE₁	G	11	G	W	S	F	G	75	D	100	G	T	H	233
NRPS-1-TE	G	11	G	W	S	L	G	75	D	100	G	D	H	231
NRPS-3-TE	G	11	G	W	S	F	G	74	D	99	G	D	H	224
NRPS-5-TE	G	11	G	H	C	V	G	76	G	101	H	A	G	220
Hybrid-2-TE₂	G	27	G	H	S	M	G	92	S	117	G	G	H	224
Hybrid-4-TE₂	G	11	G	H	S	M	G	77	S	102	G	D	H	208
Hybrid-5-TE₂	G	60	G	H	S	L	G	125	S	150	G	D	H	257

Figure 4-24: Consensus sequences of the TE domains (highlighted in yellow) from the PKS/NRPS gene loci-2, 4 and 5 and the NRPS gene loci-1, 3 and 4. The catalytic active site residues are highlighted in green. The glycine typical for the hydrolases group of TE is highlighted in blue. The putative inactive TE domain is highlighted in grey. The consensus sequence in the TE₂ domains is highlighted in pink. Conserved sequence from the literature (Konz and Marahiel, 1999) is highlighted in bold. The numbers indicate the position of amino acids. Suffices indicate the number of TE domains.

In addition to the PKS and NRPS genes, the PKS/NRPS gene locus-2 harbors a gene encoding for a glycosyltransferase (Table 4-5). This glycosyltransferase is likely to be involved in the addition of a sugar moiety to the lipopeptide resulting in a putative glycolipopeptide (Figure 4-25). Since the sugar moiety and the starter unit cannot be predicted from the sequences, the expected molecular mass of the biosynthetic product cannot be calculated. Nevertheless, HPG as an unusual building block could used to trace this metabolite in a *H. aurantiacus* DSM 785 culture.

(A)

(B)
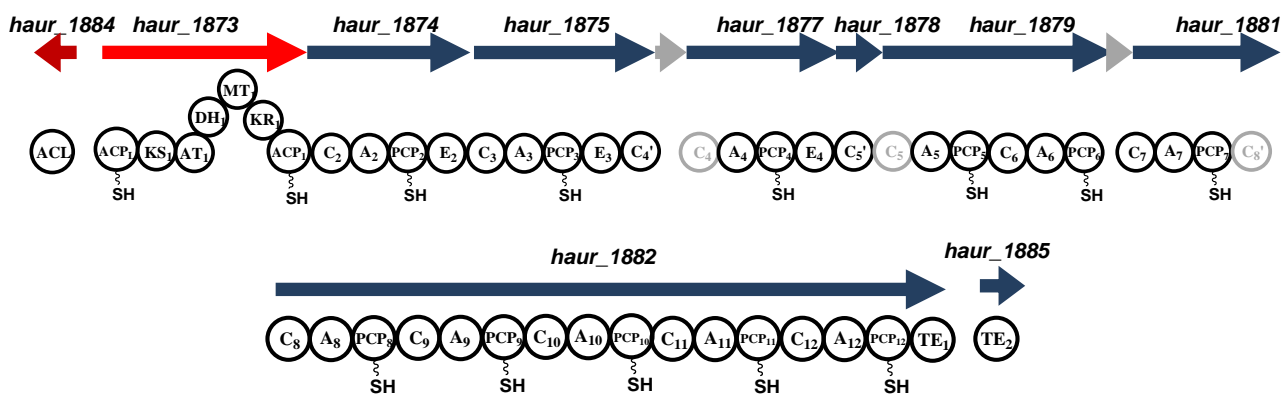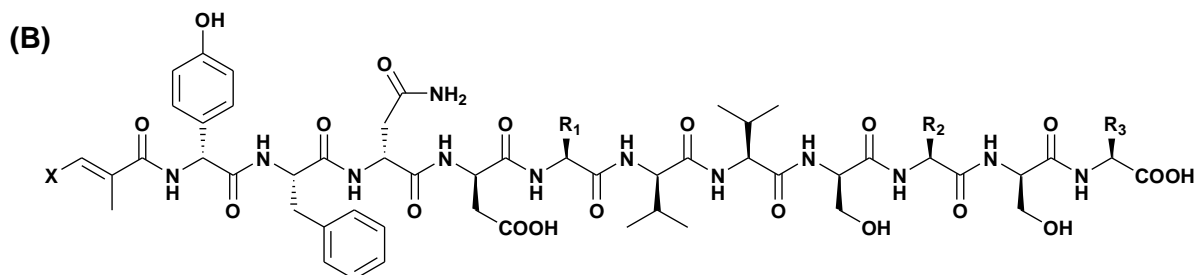
Figure 4-25: PKS/NRPS gene locus-2

(A) Putative PKS (red) and NRPS (blue) genes in PKS/NRPS gene locus-2. Brown-acyl-CoA synthetase; grey- accessory genes.

(B) Predicted strcture of the putative glycosylated lipopeptide encoded by PKS/NRPS gene locus-2. Glycosyl group to be added and the position of glycosylation is unknown. X-Unknown fatty acid side chain; R_1, R_2 and R_3-Unknown amino acid.

4.1.4.1.3 PKS/NRPS gene locus-3

The PKS/NRPS gene locus-3 is only 12.3 Kb in size and contains PKS and NRPS genes in a single ORF (Table 4-6). The biosynthesis is likely to commence with the loading of a long chain unusual starter unit on to the ACP_L by the AMP-dependent synthetase and ligase encoded by *haur_2008* (Figure 4-19). The substrate of the AT_1 domain could not be predicted from the observed consensus motif (Figure 4-20A). The substrate of A_2 domain could not predicted. Altogether, the substrates for both the PKS and NRPS modules of the PKS/NRPS gene locus-3 remain obscure. However, from the sequences all the domains i.e KS, AT, ACP, C, A, PCP in the PKS/NRPS gene locus-3 are active in the biosynthesis of the putative compound (Figure 4-20, Figure 4-26). The PKS/NRPS gene locus-3 lacks in a TE domain, instead the gene locus terminates with a C domain, i.e C_3. It is evident from the sequence alignment that this terminal C domain could be functional. Therefore, the PKS/NRPS gene locus-3 is speculated to function in conjunction with one of the other loci, i.e is as such not a complete biosynthetic gene cluster.

Gene locus	Modules	Putative functions
Haur_2008	Module-L	AMP-ACP_L
	Module-1	KS_1-AT_1-KR_1-ACP_1
	Module-2	C_2-A_2-PCP_2-C_3

Table 4-6: List of genes in PKS/NRPS gene locus-3 and their putative functions. Suffices indicate the respective modules. L, Loading module; AMP, AMP dependent synthetase and ligase; ACP, Acyl Carrier Protein; KS, Ketosynthase; AT, Acyltransferase; C, Condesation; A, Adenylation; PCP, Peptidyl Carrier Protein.
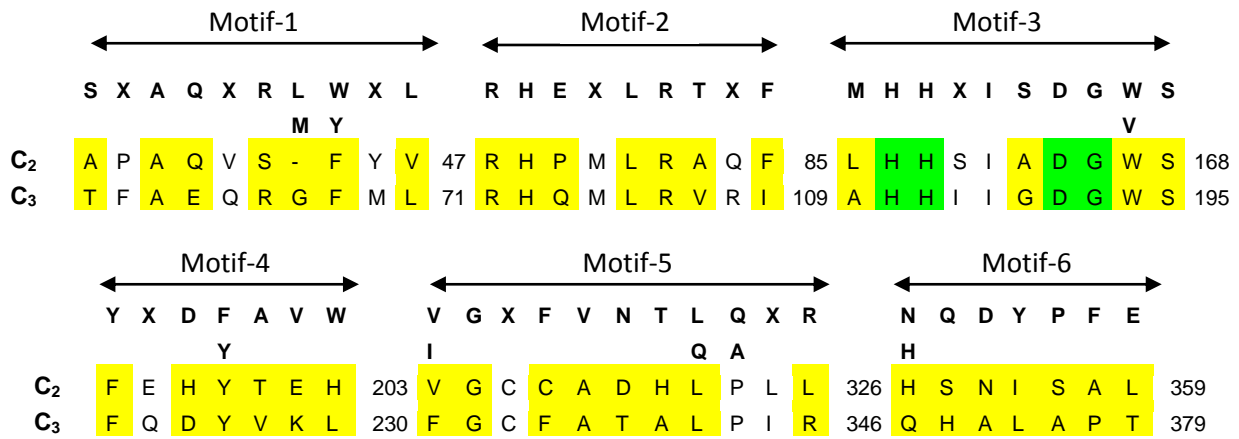
(A) C domains

(B) A domain
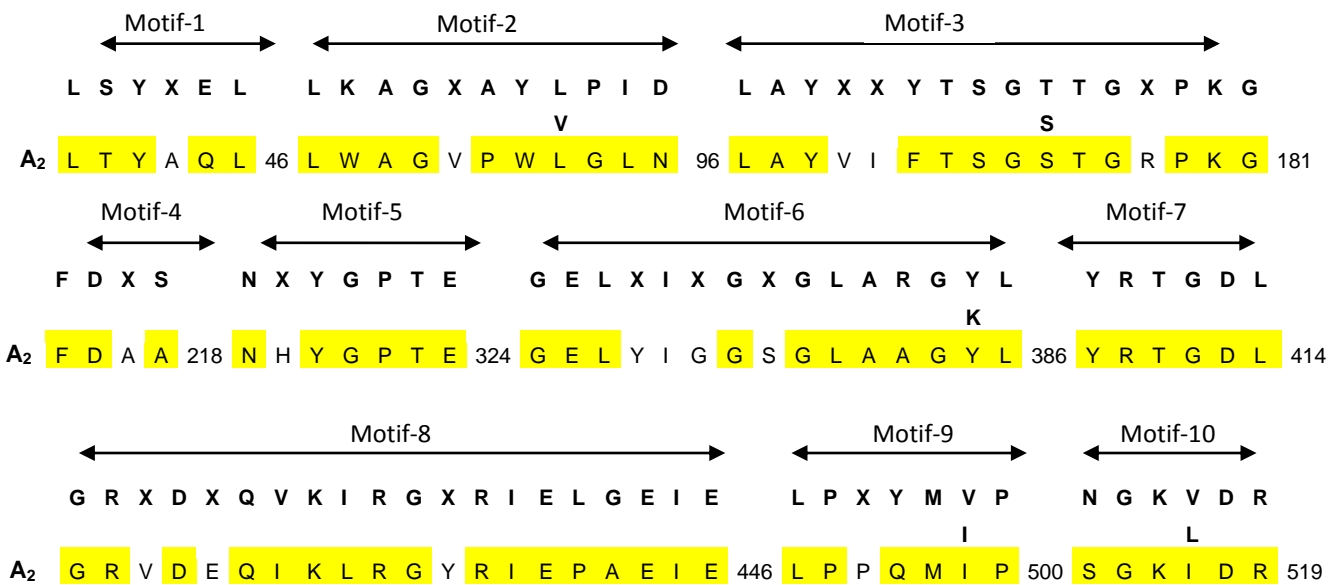
(C) PCP domain

Figure 4-26: Characteristic motifs in C (A) , A (B), PCP (C) domains of PKS/NRPS gene locus-3 (highlighted in yellow). The catalytically active residues are highlighted in green. Conserved sequences from the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers indicate the position of amino acids. Suffices indicate the respective modules.
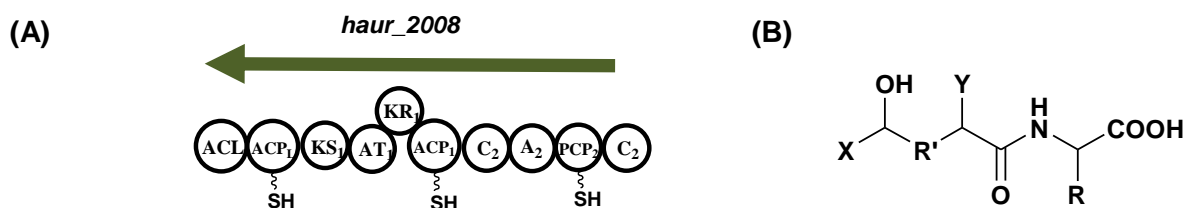
Figure 4-27: PKS/NRPS gene locus-3.

(A) Putative PKS and NRPS hybrid gene along with acyl-CoA synthetase (green) in PKS/NRPS gene locus-3.

(B) Predicted strcture of the putative lipopeptide encoded by PKS/NRPS gene locus-3. X, Unknown fatty acid side chain; R', Unknown PKS extender unit; Y, H/CH$_3$/OH; R, Unknown amino acid.

4.1.4.1.4 PKS/NRPS gene locus-4

The PKS/NRPS gene locus-4 having a size of 34 Kb (Table 4-7) is contemplated to commence the biosynthesis with loading of an unusual long chain starter unit on to the ACP$_L$, Haur_2412 by the *trans* acting AMP-dependent synthetase and ligase, Haur_2407 (Figure-4.19). The biosynthesis includes two PKS and three NRPS modules. Due to the deviations in the conserved The AT domains, i.e AT$_1$ and AT$_5$ of the PKS/NRPS gene locus-4 include the consensus sequence GHSLG (Figure 4-20A). The other consensus motif IAFH in AT1 and VAAH in AT$_5$ showed that these AT domains activate either malonyl or methylmalonyl-CoA. Sequence alignment of the PKS modules of the PKS/NRPS gene locus-4 showed that all the domains constituting the PKS region are active (Figure 4-20).

Gene locus	Modules	Putative functions
Haur_2406		Phenylalanine/Histidine ammonia-lyase
Haur_2407		AMP-dependant synthetase and ligase
Haur_2408		Glycosyltransferase
Haur_2409		ABC 2 type transporter
Haur_2410		ABC transporter related proteins
Haur_2411		Hypothetical protein
Haur_2412	Module-L	ACP$_L$
	Module-1	KS$_1$-AT$_1$-DH$_1$-KR$_1$-ACP$_1$
Haur_2413	Module-2	C$_2$-A$_2$-PCP$_2$-C$_3$
Haur_2414	Module-3	A$_3$'-A$_3$-KR$_3$-PCP$_3$

Haur_2415	Module-4	C_4-A_4-PCP_4
Haur_2416	Module-5	KS_5-AT_5-PCP_5-TE_1
Haur_2417		TE_2
Haur_2418		Glycosyltransferase
Haur_2419		O-methyltransferase

Table 4-7: List of genes in PKS/NRPS gene locus-4 and their putative functions. Suffices indicate the respective modules. L, Loading module; ACP, Acyl Carrier Protein; KS, Ketosynthase; AT, Acyltransferase; DH, Dehydratase; C, Condesation; A, Adenylation; PCP, Peptidyl Carrier Protein; TE, thioesterase.

The NRPS region of the PKS/NRPS gene locus-4 includes four A domains (Table 4-7), i.e A_2-A_4, and $A_{3'}$. The NRPS module-3 contains tandem A domains, i.e A_3 and $A_{3'}$. The amino acid sequences encoding the $A_{3'}$ domain deviates from the conserved motifs required for the binding of ATP (Figure 4-28A). Therefore, the $A_{3'}$ domain is presumed to be inactive. The substrate specificities of the other A domains from the hybrid gene locus-4 are (Table 4-7) shown in the Figure 4-28B. The NRPS module-3 also includes an active KR_3 domain. The unusual presence of KR_3 domain in the NRPS module-3 is expected to reduce the substrate loaded on the PCP_3 domain. But the Stachelhaus code was 80% identical with that of genes encoding for A domains activating α-hydroxy-isocaproic acid (Figure 4-28B). Further processing of α-hydroxy-isocaproic acid is required for the incorporation of this extender unit into the putative product. The active KR_3 domain may be expected to remain inert or act on the starter unit of the biosynthetic pathway. The C domains, i.e C_2, C_3, and C_4 of the PKS/NRPS gene locus-4 contain the six conserved motifs and the active site residues necessary for their activity (Figure 4-28A). Sequence alignment of the PCP domains, i.e PCP_2-PCP_4 showed the serine residue required for the binding of PPant arm by the action of PPTase (Figure 4-29B).

(A)

	Motif-1	Motif-2	Motif-3	Motif-4	Motif-5

	L S Y X E L	L K A G X A Y L P I D	L A Y X X Y T S G T T G X P K G	F D X S	N X Y G P T E
		V	S		

A2	L S Y A E L	52	L K A G G A Y L P I D	102	L A Y M I Y T S G S T G T P K G	181	F D F S	219	N M Y G I T E	321
A3'	V S Y H E L	36	L L L G A R Y A L L S	86	L I G L S L A S N P S G L I E A	158	- D F S	194	N Y Y G F P A	282
A3	Q S Y A A L	104	M L G G F T A V P L A	154	S A L L L F T S G S T G L P K G	244	G G I V	283	P A Y G M S E	390
A4	L S Y H E L	51	L K S G G A Y I P L D	101	L A Y I V Y T S G S T G Q P K G	181	F D A A	219	N A Y G P S E	309

	Motif-6	Motif-7	Motif-8	Motif-9	Motif-10

	G E L X I X G X G L A R G Y L	Y R T G D L	G R X D X Q V K I R G X R I E L G E I E	L P X Y M V P	N G K V D R
	K			I	L

A2	G E L Y V A G A G V F A G - F	341	Y Q T R Y L	369	G S L D S T I E L A A T P I L L Q E V E	400	L D C Q L A N	456	D G Q L D R	476
A3'	G R L Q I Q G P T I T A G Y Y	475	- - - - - D	496	G R E K D V I I I N G I N Y H N H E I E	527	I G I N P S Y	593	I G K I Q R	613
A3	G E L Y V G G A G L A R G Y W	383	Y K T G D L	409	G R R D N Q V K I R G F R I E L G E I Q	442	L P E Y M V P	499	N G K I D Y	518
A4	G E I Y V G G V G V A R G Y L	368	Y R T G D L	396	G R I D Q Q I K L R G H R I E L G E I S	429	- - - S Q A P	486	K Q K L P A	505

(B)

Gene locus	A domain	Stachelhaus code	Predicted substrates
haur_2413	A2	DFWNIGMVHK	Threonine
haur_2414	A3	GIFWLGASG*	α-OH-isocaproic acid
haur_2415	A4	GDASTIAAVC*	Tyrosine

Figure 4-28: PKS/NRPS gene locus-4 A domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS A domains (highlighted in yellow) from PKS/NRPS gene locus-4. The ten core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. The inactive A3' domain is highlighted in grey. Suffices indicate the respective module.

(B) The table lists the Stachelhaus codes extracted from the sequences of PKS/NRPS gene locus-4 A domains and their putative substrates. The asterisk indicates the unknown amino acid.

Figure 4-29: PKS/NRPS gene locus-4 C and PCP domain characteristic motifs.

(A) CLUSTALW alignment of the conserved core motifs of NRPS C domains (highlighted in yellow) from PKS/NRPS gene locus-4. The six core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules

(B) CLUSTALW alignment of the conserved core motif of NRPS PCP domains (highlighted in yellow) from PKS/NRPS gene locus-4. The core motif sequence as reported in the literature (Konz and Marahiel, 1999) is highlighted in bold. The PPant attaching serine is highlighted in green.

The PKS/NRPS gene locus-4 terminates with tandem TE domains (Table 4-7), i.e TE_1 and TE_2. The TE_1 domain contains the catalytic triad S-89/D-116/H-264 required

for activity (Figure 4-24). As described in PKS/NRPS gene locus-2, the TE_2 domain contains serine instead of aspartate in the catalytic triad, i.e S-89/S-116/H-264 (Figure 4-24). This exchange of serine residue instead of aspartate at position 116 is significant concerning the role of the second TE domain, i.e TE_2 in the biosynthesis. In addition to the PKS and NRPS genes, the PKS/NRPS gene locus-4 includes two genes *haur_2408* and *haur_2418* encoding for a glycosyltransferase for glycosylating the lipopeptide (Table 4-7). The PKS/NRPS gene locus-4 contains also a *O*-MT (*haur_2419*) for methylating a hydroxyl group of the intermediate (Table 4-7). But the exact position of *O*-methylation and glycosylation remains elusive.

The putative glycosylated lipopeptide biosynthesized by enzymes encoded by this hybrid gene locus-4 comprises of threonine and tyrosine in the peptide core (Figure 4-29). In addition the processing steps required for the incorporation of α-hydroxy isocaproic remains unclear.
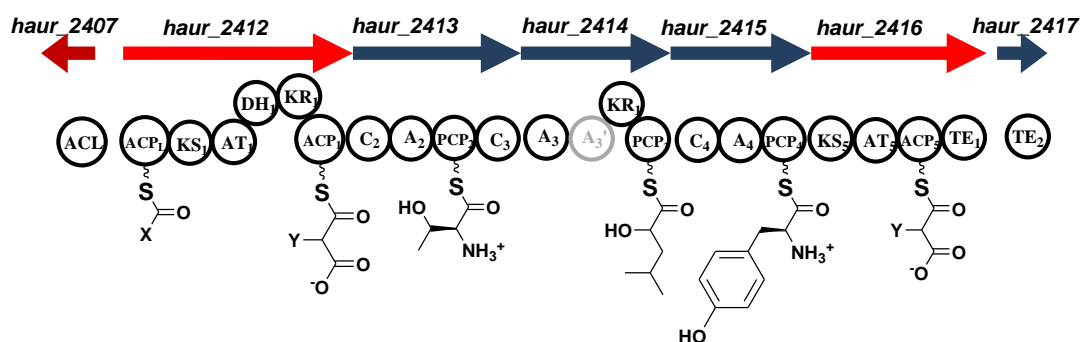
Figure 4-30: Putative PKS (red) and NRPS (blue) genes in PKS/NRPS gene locus-4 along with their respective predicted substrates. Brown, acyl-CoA synthetase. Inactive domain is highlighted in grey. X, Unknown fatty acid; Y, H/CH_3. Glycosylating groups and the positions of glycosylation is unknown. Inactive domains are highlighted in grey.

4.1.4.1.5 PKS/NRPS gene locus-5

The PKS/NRPS gene locus-5 spanning over 109 Kb is the biggest of all the biosynthetic gene loci (Table 4-8) identified from *H. aurantiacus* DSM 785. The locus comprises 16 PKS and seven NRPS modules. Biosynthesis is speculated to commence with the loading of an unknown long chain starter unit on to ACP_L by the AMP-dependent synthetase and ligase, Haur_3958 (Figure 4-19). Among the 16 PKS modules two are completely reducing modules, whereas the others are partially reducing modules with only KR domain.

| Gene locus	Modules	Putative functions
haur_3958	Module-L	$AMP\text{-}PCP_L$
haur_3959	Module-1	$KS_1\text{-}AT_1\text{-}DH_1\text{-}ER_1\text{-}KR_1\text{-}ACP_1$
haur_3960	Module-2	$KS_2\text{-}AT_2\text{-}DH_2\text{-}ER_2\text{-}KR_2\text{-}ACP_2$
haur_3961	Module-3 Module-4	$KS_3\text{-}AT_3\text{-}KR_3\text{-}ACP_3$ $KS_4\text{-}AT_4\text{-}KR_4\text{-}ACP_4$
haur_3962	Module-5	$KS_5\text{-}AT_5\text{-}KR_5\text{-}ACP_5$
haur_3963	Module-6 Module-7	$KS_6\text{-}AT_6\text{-}KR_6\text{-}ACP_6$ $KS_7\text{-}AT_7\text{-}KR_7\text{-}ACP_7$
haur_3964	Module-8 Module-9	$KS_8\text{-}AT_8\text{-}KR_8\text{-}ACP_8$ $KS_9\text{-}AT_9\text{-}KR_9\text{-}ACP_9$
haur_3965	Module-10 Module-11 Module-12	$KS_{10}\text{-}AT_{10}\text{-}KR_{10}\text{-}ACP_{10}$ $KS_{11}\text{-}AT_{11}\text{-}KR_{11}\text{-}ACP_{11}$ $KS_{12}\text{-}AT_{12}\text{-}KR_{12}\text{-}ACP_{12}$
haur_3966	Module-13 Module-14	$KS_{13}\text{-}AT_{13}\text{-}KR_{13}\text{-}ACP_{13}$ $KS_{14}\text{-}AT_{14}\text{-}KR_{14}\text{-}ACP_{14}$
haur_3967	Module-15 Module-16	$KS_{15}\text{-}AT_{15}\text{-}KR_{15}\text{-}ACP_{15}$ $KS_{16}\text{-}AT_{16}\text{-}KR_{16}\text{-}ACP_{16}$
haur_3968		3-hydroxyacyl CoA dehydrogenase
haur_3969		Hypothetical protein
haur_3970		Acyl-CoA dehydrogenase
haur_3971		Phosphatase
haur_3972	Module-17	$C_{17}\text{-}A_{17}\text{-}PCP_{17}$
haur_3973	Module-18 Module-19 Module-20 Module-21 Module-22 Module-23	$A_{18}\text{-}PCP_{18}$ $C_{19}\text{-}A_{19}\text{-}PCP_{19}$ $C_{20}\text{-}A_{20}\text{-}PCP_{20}$ $C_{21'}\text{-}C_{21}\text{-}A_{21}\text{-}PCP_{21}$ $C_{22}\text{-}A_{22}\text{-}PCP_{22}$ $C_{23}\text{-}A_{23}\text{-}PCP_{23}\text{-}TE_1$
haur_3974		TE_2

Table 4-8: List of genes in PKS/NRPS gene locus-5 and their putative functions. Suffices indicate the respective modules. L, Loading module; ACP, Acyl Carrier Protein; KS, Ketosynthase; AT, Acyltransferase; DH, Dehydratase; ER, Enoylreductase; C, Condesation; A, Adenylation; PCP, Peptidyl Carrier Protein; TE-Thioesterase.

Alignment of DH_1 domain showed variation in the catalytically active residues, G-41 and P-46 (Figure 4-31D). Therefore, DH_1 domain is presumed to be inactive. Due to inactive DH_1 domain, ER_1 with the conserved motif G-G-V-G-X-A-A-X-Q-A-A (Pawlik et al., 2007) is proposed to be inert. Other than this, all the other domains of the PKS modules, i.e AT, KS, ACP, KR, DH contain the consensus motifs required for their activity (Figure 4-31). Since the AT domains includes a different combination of consensus motifs than reported, the exact substrates could not be predicted. However, the gene locus comprises the genes *haur_3968, haur_3970* and *haur_3971* encoding for 3-hydroxyacyl CoA dehydrogenase, acyl-CoA dehydrogenase and a phosphotase required for the synthesis of hydroxymalonyl-CoA (Table 4-8) (Figure 4-4). Henceforth, it is expected that one or more of the AT domains of the hybrid gene locus-5 might activate hydroxymalonyl-CoA.

(A) AT domains

	Q		Q		G	H	S	*	G		R		#			H		§		$		V	
AT_1	Q	8	Q	58	G	H	S	M	G	89	R	112	V	A	S	H	190	T	221	N	241	V	246
AT_2	Q	8	Q	59	G	H	S	M	G	90	R	113	V	A	S	H	191	T	222	N	242	V	247
AT_3	L	31	Q	80	G	Y	S	L	G	111	R	134	H	A	F	H	208	N	239	H	259	V	264
AT_4	P	36	Q	85	G	Y	S	L	G	116	R	139	H	A	F	H	213	N	244	H	264	V	269
AT_5	N	35	Q	84	G	Y	S	L	G	115	R	138	H	A	F	H	212	N	243	H	263	V	268
AT_6	C	30	Q	80	G	Y	S	V	G	111	R	134	H	A	F	H	208	N	239	H	259	V	264
AT_7	C	31	Q	81	G	Y	S	V	G	112	R	135	H	A	F	H	209	N	240	H	260	V	265
AT_8	R	29	Q	79	G	Y	S	V	G	110	R	133	H	A	F	H	207	N	238	H	258	V	263
AT_9	C	31	Q	81	G	Y	S	V	G	112	R	135	H	A	F	H	209	N	240	H	260	V	265
AT_{10}	Q	29	Q	79	G	Y	S	L	G	110	R	133	H	A	F	H	207	N	238	H	258	V	263
AT_{11}	C	31	Q	81	G	Y	S	V	G	112	R	135	H	A	F	H	209	N	240	H	260	V	265
AT_{12}	C	31	Q	81	G	Y	S	V	G	112	R	135	H	A	F	H	209	N	240	H	260	V	265
AT_{13}	Q	29	Q	81	G	Y	S	L	G	112	R	135	H	A	F	H	207	N	238	H	258	V	263
AT_{14}	C	31	Q	81	G	Y	S	V	G	112	R	135	H	A	F	H	209	N	240	H	260	V	265
AT_{15}	Q	29	Q	79	G	Y	S	L	G	110	R	133	H	A	F	H	207	N	238	H	258	V	263
AT_{16}	C	31	Q	81	G	Y	S	V	G	112	R	135	H	A	F	H	209	N	240	H	260	V	265

(B) KS domains

	C				H				G	H		
KS₁	C	S	S	174	H	G	S	342	G	H	L	347
KS₂	C	S	S	174	H	G	T	308	G	H	L	347
KS₃	C	S	S	177	H	G	S	311	G	H	L	350
KS₄	C	S	T	175	H	G	T	300	G	H	L	349
KS₅	C	S	T	175	H	G	T	300	G	H	L	349
KS₆	C	S	T	175	H	G	T	300	G	H	L	349
KS₇	C	S	T	174	H	G	T	299	G	H	L	348
KS₈	C	S	T	173	H	G	T	298	G	H	L	347
KS₉	C	S	T	174	H	G	T	299	G	H	L	348
KS₁₀	C	S	T	175	H	G	T	300	G	H	L	349
KS₁₁	C	S	T	174	H	G	T	299	G	H	L	348
KS₁₂	C	S	T	174	H	G	T	299	G	H	L	348
KS₁₃	C	S	T	175	H	G	T	300	G	H	L	349
KS₁₄	C	S	T	174	H	G	T	299	G	H	L	348
KS₁₅	C	S	T	175	H	G	T	300	G	H	L	349
KS₁₆	C	S	T	174	H	G	T	299	G	H	L	348

(C) KR domains

Header: G X G X X G X X X A / A   S   Y   N

	G	X	G	X	X	G	X	X	X	A		S		Y		N	
KR₁	G	L	G	G	V	G	L	T	V	V	17	S	66	Y	146	N	151
KR₂	G	L	G	G	L	G	I	A	V	A	17	S	66	Y	146	N	151
KR₃	G	L	G	D	I	G	L	I	L	A	17	S	156	Y	166	N	171
KR₄	G	L	G	G	I	G	L	A	L	A	17	S	152	Y	164	N	168
KR₅	G	L	G	G	I	G	L	A	L	A	17	S	153	Y	165	N	169
KR₆	G	L	G	G	V	G	L	V	L	A	17	S	156	Y	166	N	171
KR₇	G	L	G	G	I	A	L	G	L	A	17	S	153	Y	165	N	169
KR₈	G	L	G	G	V	G	L	V	L	A	17	S	156	Y	166	N	171
KR₉	G	L	G	G	I	A	L	G	L	A	17	S	153	Y	165	N	169
KR₁₀	G	L	G	G	V	G	L	V	L	A	17	S	156	Y	166	N	171
KR₁₁	G	L	G	G	I	G	T	A	I	A	17	S	156	Y	166	N	171
KR₁₂	G	L	G	G	I	A	L	G	L	A	17	S	153	Y	165	N	169
KR₁₃	G	L	G	G	V	G	L	V	L	A	17	S	156	Y	166	N	171
KR₁₄	G	L	G	G	I	A	L	G	L	A	17	S	153	Y	165	N	169
KR₁₅	G	L	G	G	V	G	L	I	L	A	17	S	156	Y	166	N	171
KR₁₆	G	L	G	G	I	A	L	G	L	A	17	S	153	Y	165	N	169

(D) DH domains

Header: L X X H X X X G X X X X P

	L	X	X	H	X	X	X	G	X	X	X	X	P	
DH₁	L	A	E	H	Q	V	Q	D	I	V	V	L	S	46
DH₂	L	S	D	H	C	V	E	G	M	V	V	V	P	46

(E) ER domains

Header: G G V G X A A X Q X A

	G	G	V	G	X	A	A	X	Q	X	A	
ER₁	G	G	V	G	L	A	A	I	Q	L	A	151
ER₂	G	G	V	G	L	A	A	I	Q	I	A	152

(F) ACP domains

Header: G X X S

	G	X	X	S		
ACPₗ	G	L	D	S	A	32
ACP₁	G	F	D	S	L	32
ACP₂	G	L	D	S	L	32
ACP₃	G	G	N	S	L	30
ACP₄	G	G	N	S	L	30
ACP₅	G	G	N	S	L	30
ACP₆	G	G	N	S	L	30
ACP₇	G	G	N	S	L	30
ACP₈	G	G	N	S	L	30
ACP₉	G	G	N	S	L	30
ACP₁₀	G	G	N	S	L	30
ACP₁₁	G	G	N	S	L	30
ACP₁₂	G	G	N	S	L	30
ACP₁₃	G	G	N	S	L	30
ACP₁₄	G	G	N	S	L	30
ACP₁₅	G	G	N	S	L	30
ACP₁₆	G	G	N	S	L	30

Figure 4-31: Characteristic motifs in AT (A), KS (B), KR (C), DH (D), ER (E) and ACP (F) domains of PKS modules extracted from PKS/NRPS gene locus- 5 (highlighted in yellow).

Catalytically active residues are highlighted in green. Consensus sequences from the literature (Yadav et al., 2003; Zhang et al., 2006; Bachmann and Ravel., 2009; Pawlik et al., 2007) are highlighted in bold. * indicates the invariant residues Q/M/I/L/V/F/A/M, # indicates S/F/P, § indicates T/A/N/G/E/D/S and $ indicates N/S/H/Q. The numbers represents the position of amino acids. Suffices indicate the respective modules.

The PKS/NRPS gene locus-5 includes seven NRPS modules (Table 4-8). Of these, module-18 lacks a C domain, whereas module-22 has supposedly functional tandem C domains, i.e $C_{22'}$ and C_{22}. According to bioinformatic analysis, NRPS module-18 with active A_{18} and PCP_{18} domains, has however no C_{18} domain, and thus is anticipated to be skipped during biosynthesis. The role of tandem C domains $C_{22'}$ and C_{22} in the NRPS module-22, remains obscure. Sequence alignment of all the domains of the NRPS modules, i.e C, A, PCP showed the presence of conserved motifs required for their activity (Figure 4-32, Figure 4-33). From the sequences of the A domains from the hybrid gene locus-5 the putative substrates activated by the respective A domains were identified (Figure 4-32B).

Results

(A)

Motif-1 → L S Y X E L
Motif-2 → L K A G X A Y L P I D (V)
Motif-3 → L A Y X X Y T S G T T G X P K G (S)
Motif-4 → F D X S
Motif-5 → N X Y G P T E

	Motif-1		Motif-2		Motif-3		Motif-4		Motif-5	
A_{17}	L T Y G E L	52	L K A G A A Y V P L D	103	P A Y V I Y T S G S T G Q P K G	181	F D V C	219	N S Y G V T E	361
A_{18}	V T Y W Q L	49	L K A G G A Y L P L D	100	L A Y V I Y T S G S T G R P K G	181	F D A A	218	N N Y G P T E	347
A_{19}	L T Y A E L	54	L K A G G A Y V P L D	105	L A Y I I Y T S G S T G R P K G	181	F D I S	218	N F Y G P S E	344
A_{20}	L T Y A E L	54	L K A G G A Y V P L D	105	L A Y I I Y T S G S T G R P K G	181	F D I S	218	N F Y G P S E	344
A_{21}	L T Y A E L	54	L K A G G A Y V P L D	105	L A Y I I Y T S G S T G R P K G	181	F D I S	218	N F Y G P S E	344
A_{22}	L L Y R D I	36	L K S G G A F A L V N	87	A A C L V Q T D G A A G T R R L	165	T G D A	202	N L Y G A A E	338
A_{23}	L T Y R E L	51	V I A G G A Y V P L D	102	L A Y M I Y T S G S T G R P K G	181	F D L T	218	N E Y G P T E	353

Motif-6 → G E L X I X G X G L A R G Y L (K)
Motif-7 → Y R T G D L
Motif-8 → G R X D X Q V K I R G X R I E L G E I E
Motif-9 → L P X Y M V P (I)
Motif-10 → N G K V D R (L)

	Motif-6		Motif-7		Motif-8		Motif-9		Motif-10	
A_{17}	G E L Y L G G S R L A L G Y W	382	Y R T G D A	412	G R I D Q Q V K L H G V R I E L G E I E	445	L P E H M V P	498	N G K I D R	518
A_{18}	G E L Y V G G V G L A R Y Y L	370	Y R T G D R	400	G R I D F Q V K V R G F R I E L G E I E	433	L P E Y M V P	488	N G K I D R	508
A_{19}	G E L Y L G G A G V S R G Y F	367	Y R T G D R	397	G R M D R Q V K L R G F R I E L G E V E	430	L P E Y M V P	485	S G K L D R	505
A_{20}	G E L Y L G G A G V S R G Y F	367	Y R T G D R	397	G R M D R Q V K L R G F R I E L G E V E	430	L P E Y M V P	485	S G K L D R	505
A_{21}	G E L Y L G G A G V S R G Y F	367	Y R T G D R	397	G R M D R Q V K L R G F R I E L G E V E	430	L P E Y M V P	485	S G K L D R	505
A_{22}	G E L Y L G G A G V S R G Y F	360	Y R T G D R	390	G R V E Q Q V K L R G F R I E L G E V E	423	L P E Y M V P	478	S G K L D R	498
A_{23}	G E L Y I G G A G V G R G Y H	376	Y K T G D L	406	G R V D H Q V K V R G F R I E L E E I E	439	L P A Y M A P	494	N G K V D R	514

103

(B)

Gene locus	A domain	Stachelhaus code	Predicted substrate
haur_3972	A_{17}	DVGEIGSIDK	Ornithine
haur_3973	A_{18}	DAWFLGNVVK	Leucine
haur_3973	A_{19}	DIFNNAFTYK	Alanine
haur_3973	A_{20}	DIFNNAFTYK	Alanine
haur_3973	A_{21}	DIFNNAFTYK	Alanine
haur_3973	A_{22}	GDWRLALLDK	Unknown
haur_3973	A_{23}	DLTKVGEVGK	Asparagine

Figure 4-32: PKS/NRPS gene locus-5 A domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS A domains (highlighted in yellow) from PKS/NRPS gene locus-5. The ten core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules.

(B) The table lists the Stachelhaus codes extracted from the sequences of PKS/NRPS gene locus-5 A domains and their putative substrates

(A) C domains

	Motif-1											Motif-2									Motif-3									
	S	X	A	Q	X	R	L	W	X	L		R	H	E	X	L	R	T	X	F	M	H	H	X	I	S	D	G	W	S
							M		Y																					V
C_{17}	S	F	S	Q	E	R	L	W	F	L	16	R	H	E	I	L	R	T	S	F	61	L	H	H	I	A	Y	D	E	W	S	146
C_{19}	S	F	A	Q	E	R	L	W	F	L	16	R	H	A	P	L	R	S	T	F	61	I	H	H	I	I	F	D	G	W	S	144
C_{20}	S	F	A	Q	E	R	L	W	F	L	16	R	H	A	P	L	R	S	T	F	61	I	H	H	I	I	F	D	G	W	S	144
$C_{21'}$	S	V	D	Q	E	R	F	W	F	V	16	R	H	E	S	L	R	A	F	F	61	L	H	H	S	V	F	D	G	W	S	144
$C_{21'}$	S	F	A	Q	E	R	L	W	F	L	16	R	H	E	I	L	R	T	T	F	61	I	H	H	I	V	F	D	G	W	S	144
C_{22}	S	F	A	Q	E	R	L	W	F	L	16	R	H	E	S	L	R	T	T	F	61	L	H	H	S	I	S	D	G	W	S	144
C_{23}	S	F	A	Q	E	R	L	W	F	L	16	R	H	E	S	L	R	T	T	F	61	L	H	H	S	I	S	D	G	W	S	144

	Motif-4							Motif-5												Motif-6							
	Y	X	D	F	A	V	W	V	G	X	F	V	N	T	L	Q	X	R		N	Q	D	Y	P	F	E	
				Y						I						Q	A					H						
C_{17}	Y	G	D	Y	A	Y	W	182	I	G	F	F	V	N	S	L	P	L	R	301	H	Q	D	V	P	I	E	333
C_{19}	Y	A	D	Y	A	R	W	180	I	G	M	F	V	N	T	L	A	L	R	299	H	Q	D	L	P	F	E	331
C_{20}	Y	A	D	Y	A	R	W	180	I	G	M	F	V	N	T	L	A	L	R	299	H	Q	D	L	P	F	E	331
$C_{21'}$	Y	A	D	Y	A	A	W	183	I	G	L	F	I	N	S	L	P	L	R	302	H	Q	N	I	P	L	T	334
$C_{21'}$	Y	A	D	Y	A	R	W	183	I	G	M	F	V	N	T	L	A	L	R	302	H	Q	D	L	P	F	E	334
C_{22}	Y	A	D	Y	A	L	W	183	I	G	M	F	V	N	T	L	A	L	R	302	H	Q	D	V	P	F	E	334
C_{23}	Y	A	D	Y	A	L	W	183	I	G	M	F	V	N	T	L	A	L	R	302	H	Q	D	V	P	F	E	334

(B) PCP domains

	D	X	F	F	X	L	G	G	H	S	L	
										D	I	
PCP₁₇	D	N	F	F	E	L	G	G	H	S	L	30
PCP₁₈	S	N	F	F	D	L	G	G	H	S	L	30
PCP₁₉	S	N	F	F	D	L	G	G	H	S	L	32
PCP₂₀	S	N	F	F	D	L	G	G	H	S	L	32
PCP₂₁'	S	N	F	F	A	L	G	G	H	S	L	32
PCP₂₁	S	N	F	F	A	L	G	G	H	S	L	32
PCP₂₂	Q	N	F	F	D	L	G	G	H	S	L	30

Figure 4-33: PKS/NRPS gene locus-5 C and PCP domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS C domains (highlighted in yellow) from PKS/NRPS gene locus-5. The six core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules.

(B) CLUSTALW alignment of the conserved core motif of NRPS PCP domains (highlighted in yellow) from PKS/NRPS gene locus-5. The core motif sequence as reported in the literature (Konz and Marahiel, 1999) is highlighted in bold. The PPant attaching serine is highlighted in green color.

From the sequences, the PKS/NRPS gene locus-5 is predicted to encode the biosynthetic enzymes for a lipopeptide (Figure 4-34). The peptide core of the putative compound comprises eight amino acids including L-ornithine, L-leucine, three times L-alanine and L-asparagine (Figure 4-33B). One of the A domain Stachelhaus codes showed only 50% homology to that of L-serine. Therefore, this domain is suspected to activate some other substrate. Considering the polyketide backbone, the conserved residues allowing for the prediction of substrate selectivity remain obscure. Since, biosynthetic genes encoding for the formation of hydroxymalonyl-CoA were identified for at least one of the PKS modules, this is expected to load this unit. Nevertheless, it would be interesting to isolate the compound because of the unusual motifs in the AT domains.

(A)
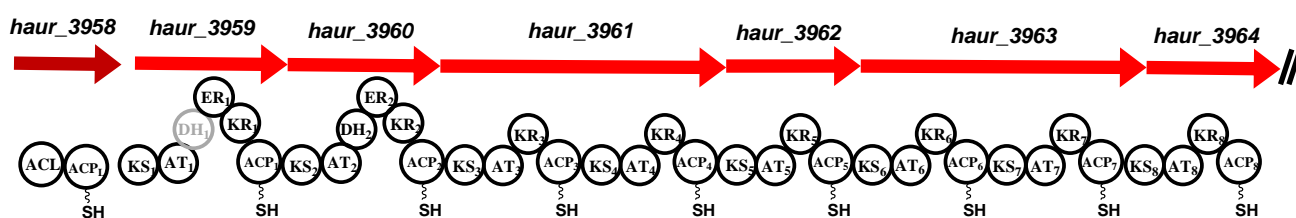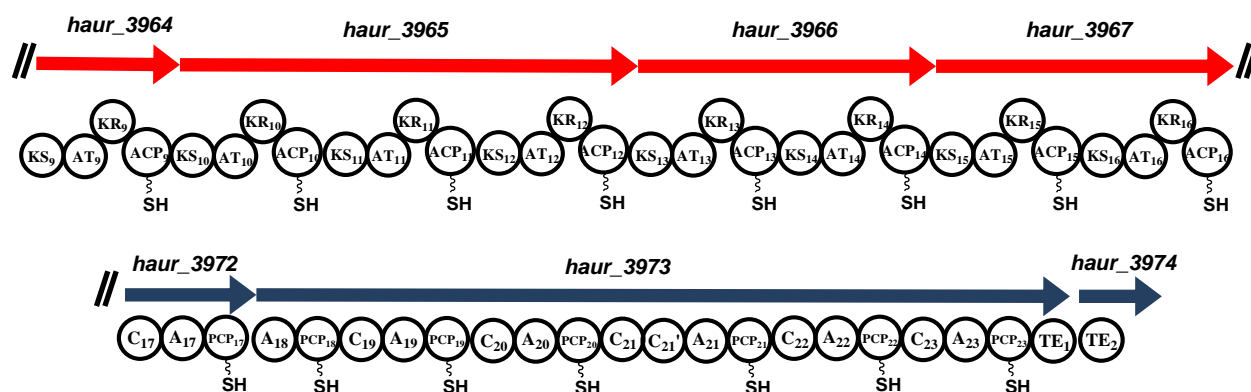
(B)
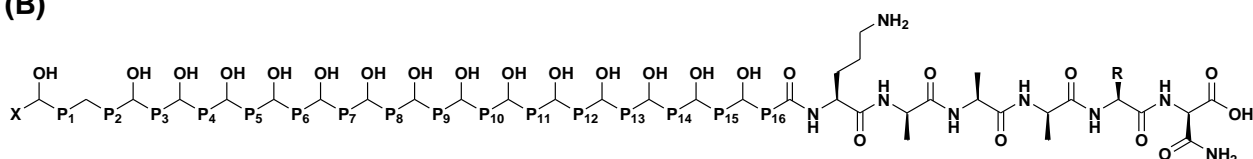
Figure 4-34: PKS/NRPS gene locus-5.

(A) Putative PKS (red) and NRPS (blue) genes in PKS/NRPS gene locus-5. Brown-acyl-CoA synthetase.

(B)Predicted strcture of the putative lipopeptide encoded by PKS/NRPS gene locus-5. X-Unknown fatty acid side chain; P_{1-16}-Unknown PKS extender units; R-Unknown amino acid.

4.1.4.1.6 NRPS gene locus-1

The NRPS gene locus-1 comprising 8.6 Kb is rather a small biosynthetic gene locus with only three NRPS modules, including the loading module (Table 4-9). The amino acid sequence of the C_2 domain is very short and consequently the conserved active motifs could not be identified from the sequence (Figure 4-37A). Moreover, the A_2 domain lacks in the core motifs-6 to10 (Figure 4-37A). Therefore, the second NRPS module is speculated to be skipped during the biosynthesis of the dipeptide. Sequence alignment of the A domains showed that the A_L domain activates glutamine and the A_1 domain activates HPG (Figure 4-37B). The condensed dipeptide is then most likely to be transferred to the PCP3 which would subsequently released by the action of the TE domain of *haur_1573* (Figure 4-35)*.

Gene locus	Modules	Putative functions
haur_1574	Module-L Module-1	A_L-PCP$_L$ C_1-A_1-PCP$_1$
haur_1572	Module-2	C_2 -A_2

haur_1573		PCP₂-TE

Table 4-9: List of genes in NRPS gene locus-1 and their putative functions. L, Loading module; A, Adenylation; C, Condensation; PCP, Peptidyl Carrier Protein; TE, Thioesterase.
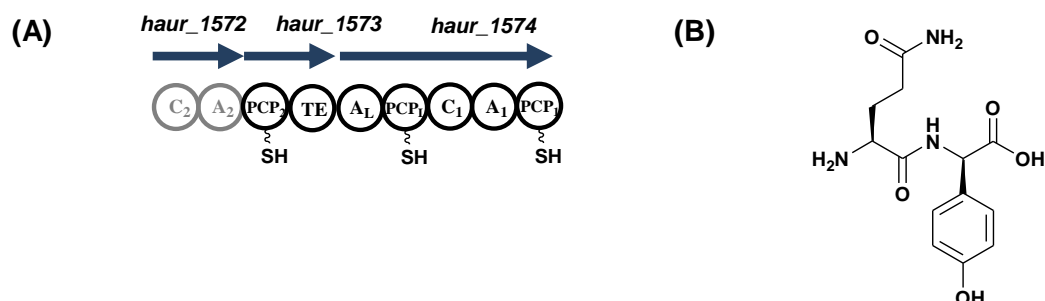
Figure 4-35: NRPS gene locus-1.

(A) Putative NRPS (blue) genes in NRPS gene locus-1. Inactive domains are highlighted in grey.

(B)Predicted strcture of the putative dipeptide encoded by NRPS gene locus-1.

(A) C domains
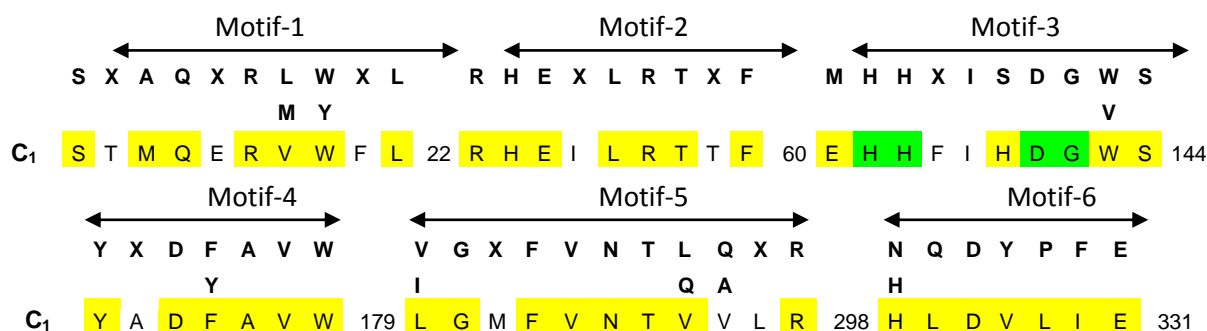
(B) PCP domains
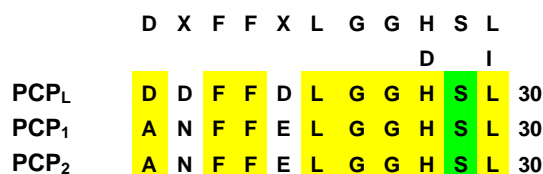
Figure 4-36: Characteristic motifs in C (A) and PCP (B) domains of NRPS gene locus-1 (highlighted in yellow). The catalytically active site residues are highlighted in green. Conserved sequences from the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers indicate the position of amino acids. Suffices indicate the respective modules.

(A)

(B)
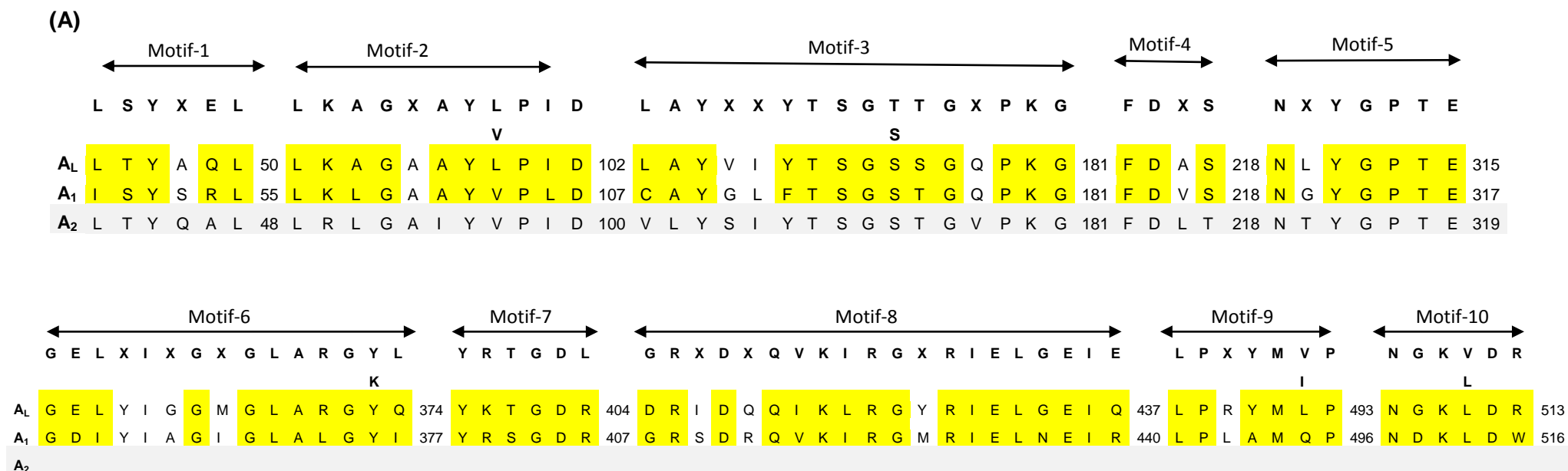
Gene locus	A domain	Stachelhaus code	Predicted substrates
haur_1574	A$_L$	DAWQFGLIDK	Glutamine
haur_1574	A1	DVYYLGGICK	HPG

Figure 4-37: NRPS gene locus-1 A domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS A domains (highlighted in yellow) from NRPS gene locus-1. The ten core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. Sufices indicate the respective modules. The numbers represent the amino acid positions in the respective sequences.

(B) The table lists the Stachelhaus codes extracted from the sequences of NRPS gene locus-1 A domains and their putative substrates.

4.1.4.1.7 NRPS gene locus-2

NRPS gene locus-2 of 14 Kb size is homologous to the myxochelin biosynthetic gene cluster. The locus includes genes, i.e encoding isochorismate synthase (*haur_1802*), isochorismatase (*haur_1804*) and 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (*haur_1800*) required for the synthesis of 2,3- dihydroxybenzoic acid (DHBA) (Table 4-10) (Figure 4-38A). These regions are also bordered by genes encoding for iron transport and utilization proteins. Therefore, it is anticipated that this gene locus encodes for a siderophore. Based on the Stachelhaus code extracted from the core motifs of the A domain amino acid sequence, the A domain is expected to activate lysine (Figure 4-39). This activation is consistent with that during myxochelin biosynthesis. Comparing the biosynthetic gene cluster for myxochelin from *M. xanthus* DK1622 with NRPS gene locus-2, that latter was identified to encode for myxochelin A biosynthesis (Figure 4-38). Myxochelin A and B are the siderophores with two units of 2,3-dihydroxybenzoic acid (DHBA) condensed with the amino acid lysine. Till date, the myxochelin siderophore is known only from Myxobacteria (Gaitatzis et al., 2001; Silakowski et al., 2000). Sequence alignments of the NRPS C, A, PCP and R domains showed that all the domains are active and likely to participate in the biosynthesis of myxochelin A (Figure 4-38, Figure 4-39, Figure 4-40, Figure 4-41).

Gene locus	Putative functions
Haur_1800	2,3-dihydro-DHBA dehydrogenase
Haur_1801	DAHP synthase
Haur_1802	Isochorismate synthase
Haur_1803	DHBA-AMP ligase
Haur_1804	Isochorismatase
Haur_1805	C-A-PCP-R
Haur_1806	Periplasmic binding protein
Haur_1807	Transport system permease
Haur_1808	Transport system permease
Haur_1809	ABC transporter

Table 4-10: List of genes in NRPS gene locus-2, DHBA, 2,3-dihydroxybenzoic acid; DAHP, 3-deoxy-D-arabinoheptulosonate 7-phosphate.
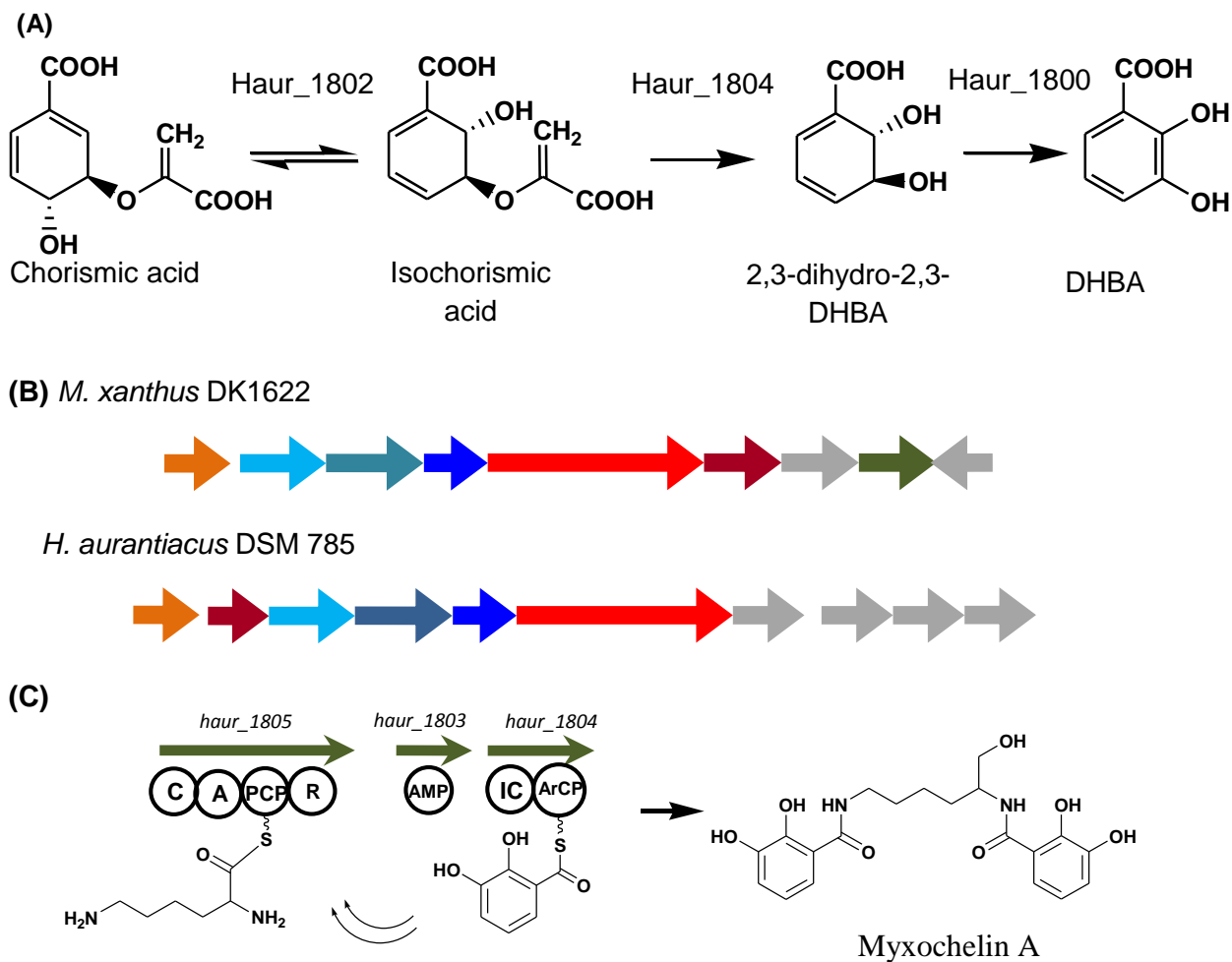
Figure 4-38: Myxochelin in *H. aurantiacus.*

(A) 2,3-dihydroxybenzoic acid (DHBA) biosynthetic pathway from *Stigmatella aurantiaca* Sg15. The gene *haur_1802* represents isochorismate synthase, *haur_1804* for Isochorismatase and Haur_1800 2,3-dihydro-DHBA dehydrogenase.

(B) Myxochelin biosynthetic gene cluster from *M. xanthus* DK1622 and *H. aurantiacus* DSM 785. Red, NRPS gene; green, aminotransferase; grey, iron transport and utilization proteins, others, genes for the biosynthesis and loading of DHBA.

(C) Proposed myxochelin biosynthetic pathway in *H. aurantiacus* DSM 785. Due to the absence of aminotransferase gene (highlighted in green) in (B) the final product of the pathway in *H. aurantiacus* DSM 785 is expected to be myxochelin A. AMP, DHBA AMP dependent synthetase and ligase; IC, Isochorismate synthase; ArCP, Aryl Carrier Protein; C, Condensation; A, Adenylation; PCP, Peptidyl Carrier Protein; R, Reductase. The two arrows indicate the iterative usage of DHBA.

Motif-1 | Motif-2 | Motif-3

L S Y X E L | L K A G X A Y L P I D | L A Y X X Y T S G T T G X P K G
                    V   S

NRPS-2 L S Y A E L 63 L C C G A A Y L P L D 113 P A Y I M Y T S G S T G E P K G 181
MxcG L T Y A A L 55 L F S G A G Y L P V D 105 L A Y V I Y T S G S T G Q P N G 181

Motif-4 | Motif-5 | Motif-6 | Motif-7

F D X S | N X Y G P T E | G E L X I X G X G L A R G Y L | Y R T G D L
                                           K

NRPS-2 F D A S 218 N T Y G P T E 319 G D L Y L L G N G L A T G Y Y 378 Y R T G D R 406
MxcG F D A S 218 N T Y G P T E 319 G E L H L L G G A L A R G Y L 384 Y R T G D K 412

Motif-8 | Motif-9 | Motif-10

R X D X Q V K I R G X R I E L G E I E | L P X Y M V P | N G K V D R
                                        I   L

NRPS-2 G R S D D E F K I S G Q R V T P A E I E 437 L P A A V I P 493 A G K I D R 512
MxcG G R V D D E F K I S G H R I D P S E V E 444 L P A P M V P 500 T G K L D R 519

Figure 4-39: Characteristic core motifs of A domains (highlighted in yellow) from the NRPS gene locus-2 in *H. aurantiacus* DSM 785, MxcG-*Myxococcus xanthus* DK1622 (Acc no-NC_008095.1). Conserved sequences from the literature are highlighted in bold. The numbers indicate the position of amino acids.

(A) C domains

Motif-1 | Motif-2 | Motif-3

S X A Q X R L W X L | R H E X L R T X F | M H H I S D G W S
                 M Y   V

NRPS-2 S A A Q H G I W L G 16 E A A T L H L R F 61 A H H I A L D G Y S 121
MxcG S A A Q H G I W V G 16 E A E A L H A R F 61 A H H I A L D G F G 122

Motif-4 | Motif-5 | Motif-6

Y X D F A V W | V G X F V N T L Q X R | N Q D Y P F E
        Y   I   Q A   H

NRPS-2 T L D R E F W 192 P C M A M N I V P L C 290 H G R Y R Y E 323
MxcG A A D H A F W 193 P C M A M N I V P L R 290 H L R Y R Y E 323

(B) PCP domains

D X F F X L G G H S L
                      D   I

NRPS-2-PCP A D F F G L G G Q S L 30
MxcG-PCP D D F F D L G G Q S L 30

Figure 4-40: Characteristic motifs in the C (A) and PCP (B) domains (highlighted in yellow) identified from the NRPS gene locus-2 in *H. aurantiacus* DSM 785. MxcG is from *M. xanthus* DK1622 (Acc no-NC_008095.1). The catalytically active residues are highlighted in green. The PPant attaching serine in PCP domain is highlighted in green. The conserved residues from the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers indicate the position of amino acids.
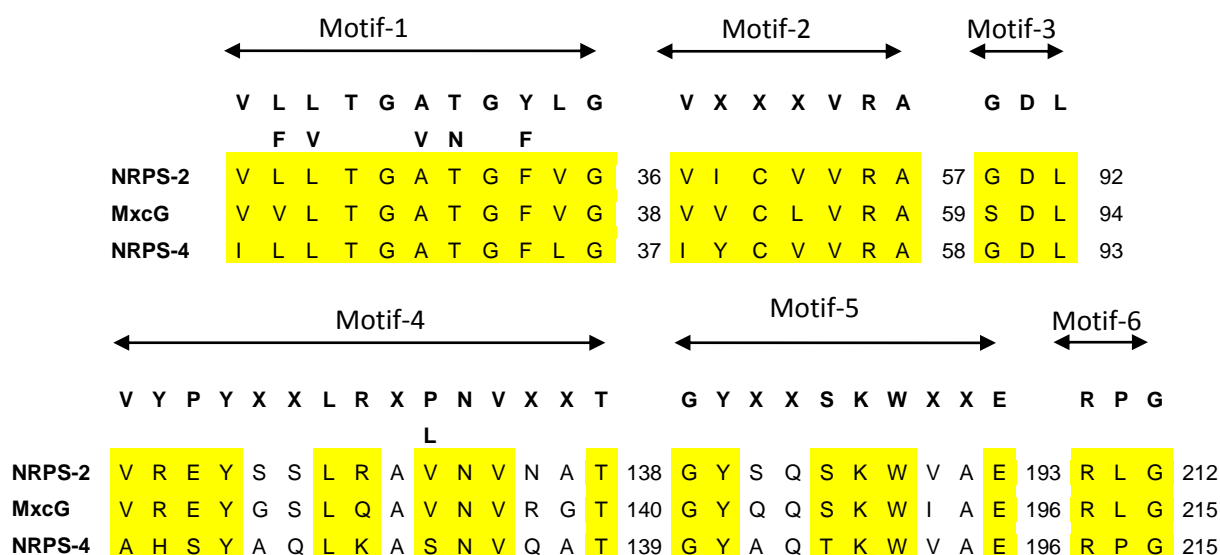

Figure 4-41: Conserved motifs in the R domain (highlighted in yellow) from the NRPS gene loci-2 and 4 in *H. aurantiacus* DSM 785, MxcG-R-R domain from *Myxococcus xanthus* DK1622 (Acc no-NC_008095.1). Conserved residues from the literature (Konz and Marahiel, 1999) are highlighted in yellow. The numbers indicate the position of amino acids.

4.1.4.1.8 NRPS gene locus-3

The NRPS gene locus-3, spanning 16 Kb comprises five NRPS modules including the loading module (Table 4-11). The module-1 without A_1 domain is anticipated either to be skipped in the biosynthesis or involved making use of adjacent A domains. The substrate specificities of the A domains are listed in Figure 4-43B. Since the substrates encoding Stachelhaus code for A_3 and A_4 was only 50% identical to that known for serine it is expected that these A domains activate some other amino acids. The other A domains, i.e A_L and A_2 was predicted to activate alanine and threonine respectively. From the random location of the initiation, elongation and the termination modules in the NRPS gene locus-5, the biosynthesis is likely to stutter between the modules. The TE domain contains the catalytic triad S-89/D-166/H-264 required for its function (Figure 4-24). The NRPS gene locus-3

includes the superfluous $PCP_{3'}$ domain. Sequences of the $PCP_{3'}$ domain showed that it is functional (Figure 4-44B). Therefore, this domain might function in conjunction with any of the other loci, most likely with the PKS/NRPS gene locus-3 terminating with a C domain synthesizing a tetrapeptide (Figure 4-42) or just transfers the building blocks between the adjacenet modules. However, it is interesting to isolate the compound and look for the biosynthesis of the latter.

Gene locus	Modules	Putative functions
haur_2089	Module-4	C_4-A_4-PCP_4-TE
haur_2090	Module-L Module-1	A_L-PCP_L C_1-PCP_1
haur_2091	Module-2	C_2-A_2-PCP_2
haur_2092	Module-3	$PCP_{3'}$-$E_{3'}$-C_3-A_3-PCP_3

Table 4-11: List of genes in NRPS gene locus-3 and their putative functions. L, Loading module; A, Adenylation; C, Condensation; PCP, Peptidyl Carrier Protein; E-Epimerization; TE, Thioesterase.

(A)

(B)
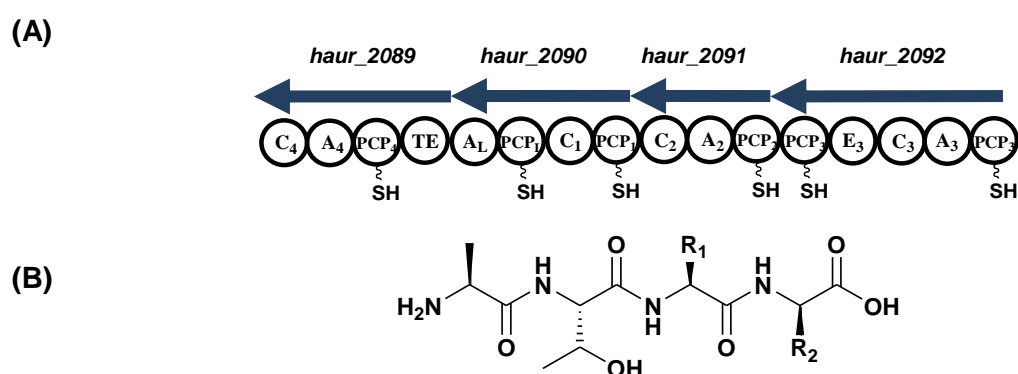
Figure 4-42: NRPS gene locus-3.

(A) Putative NRPS (blue) genes in NRPS gene locus-3.

(B)Predicted strcture of the putative tetrapeptide encoded by NRPS gene locus-3 $R_{1,2}$-Unknown amino acid.

(A)

(B)
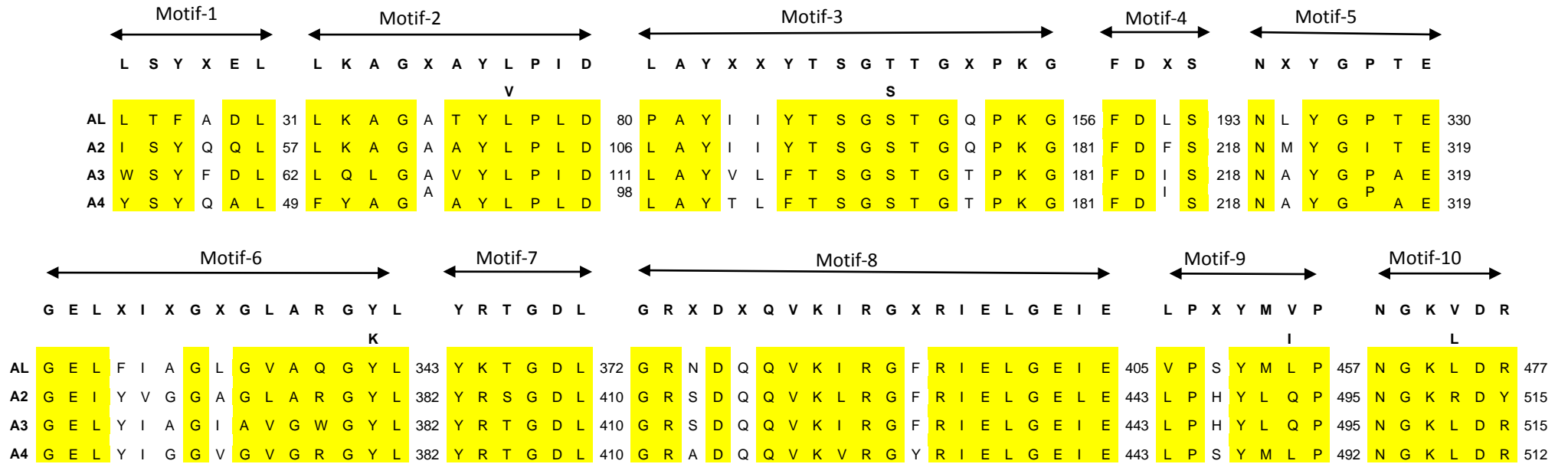
Gene locus	A domain	Stachelhaus code	Predicted substrate
haur_2090	A_L	DLFNNALTYK	Alanine
haur_2091	A_2	DFWNIGMVHK	Threonine
haur_2092	A_3	DIWELTADDK	Unknown
haur_2089	A_4	DIWELTADDK	Unknown

Figure 4-43: NRPS gene locus-3 A domains.

(A)- CLUSTALW alignment of the conserved core motifs of NRPS A domains (highlighted in yellow) from NRPS gene locus-3. The ten core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules.

(B) The table lists the Stachelhaus codes extracted from the sequences of NRPS gene locus-3 A domains and their putative substrates.

(A) C domains

(B) PCP domains

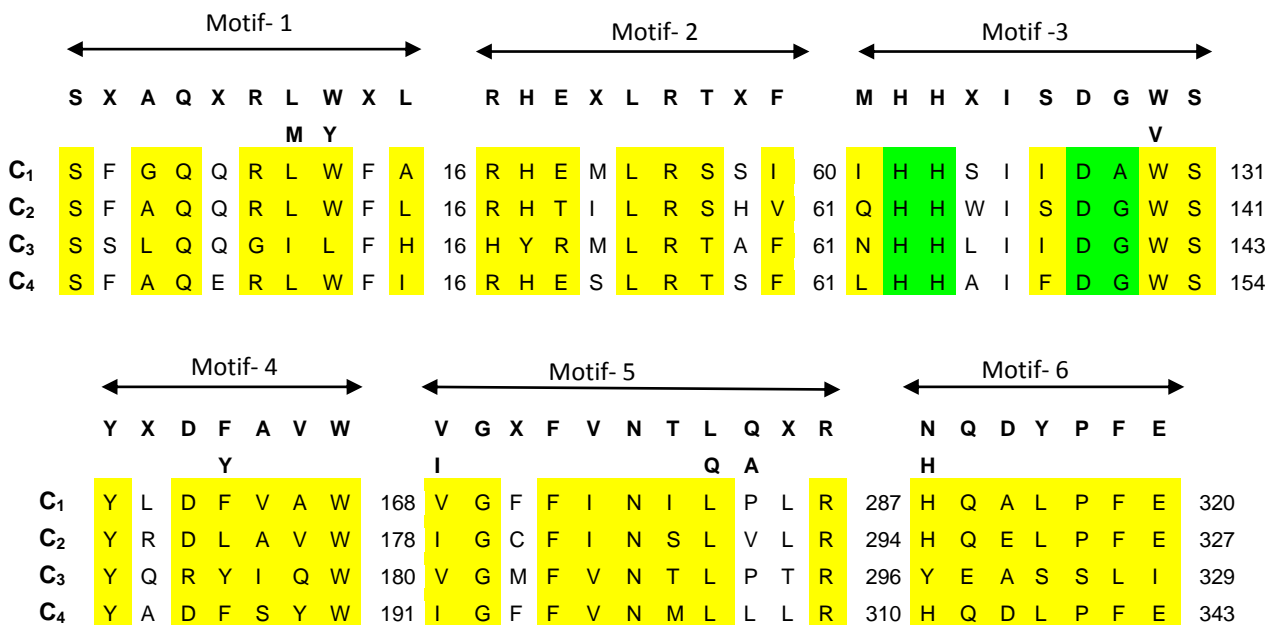**Figure 4-44:** NRPS gene locus-3 C and PCP domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS C domains (highlighted in yellow) from NRPS gene locus-3. The six core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules.

(B) CLUSTALW alignment of the conserved core motif of NRPS PCP domains (highlighted in yellow) from NRPS gene locus-3. The core motif sequence as reported in the literature (Konz and Marahiel, 1999) is highlighted in bold. The PPant attaching serine residue is highlighted in green.

4.1.4.1.9 NRPS gene locus-4

The NRPS gene locus has a size of 18.5 Kb and includes four NRPS modules, terminating with a functional R_4 domain (Table 4-12). The biosynthesis related to this gene locus begins with the loading of a long chain unknown starter unit on to the ACP_L by the AMP-dependent synthetase and ligase (Figure 4-19) encoded by *haur_2108*. The NRPS module-4 contains tandem C domains, $C_{3'}$ and C_3 in two adjacent ORFs. Alignment of C domain amino acid sequence of NRPS gene locus-4 showed deviations in the histidine active site motif in C_1, C_2 and C_3 domain (Figure 4-45A). Therefore, it is presumed that C_1, C_2 and C_3domains are inactive. Hence, these inactive C domains, i.e C_1 and C_2 might hinder the module-1 and module-2 from taking part in the biosynthesis of the peptide. However, module-3 is expected to resumes its activity by the action of $C_{3'}$ domain. The rest of the NRPS domains and ACP_L were identified to contain the respective conserved motifs (Figure 4-45, Figure 4-46). The synthesized putative lipodipeptide is expected to be reductively released by the R_4 domain (Figure 4-47). Downstream to the NRPS gene locus-4 is the *mbtH-like* gene, *haur_2109* which was already reported to be involved in the biosynthesis of non-ribosomal peptides (Boll et al., 2011, Felnagle et al., 2010).

Gene locus	Modules	Putative functions
haur_2109		MbtH-like protein
haur_2108	Module-L	AMP-ACP_L
	Module-1	C_1-A_1-PCP_1
	Module-2	C_2-A_2-PCP_2
haur_2107	Module-3	$C_{3'}$
haur_2106		C_3-A_3-PCP_3
	Module-4	C_4-A_4
haur_2105		Taurine catabolism dioxygenase (TauD/TfdA)
haur_2104	Module-4	PCP_4-R_4

Table 4-12: List of genes in NRPS gene locus-4 and their putative functions. L, Loading module; AMP, AMP-dependent synthetase and ligase; ACP, Acyl Carrier Protein; C, Condensation, A-Adenylation, PCP-Peptidyl Carrier Protein, R-Reductase.

(A) C domains

			Motif-1										Motif-2									Motif-3								
	S	X	A	Q	X	R	L	W	X	L		R	H	E	X	L	R	T	X	F		M	H	H	X	I	S	D	G	W	S
							M	Y																							V
C₁	S	F	G	E	Q	A	L	W	A	L	28	R	H	A	A	L	R	T	N	F	74	Q	H	-	T	V	V	D	L	Q	S	151
C₂	A	R	M	Q	A	G	V	I	F	H	43	R	H	A	V	L	R	T	S	F	89	F	H	D	T	I	F	D	G	W	S	172
C₃'	S	L	A	Q	Q	R	L	W	L	V	15	R	H	E	V	L	R	T	S	I	61	L	H	H	I	I	A	D	G	W	S	136
C₃	S	P	Q	Q	T	H	T	W	L	V	18	R	H	E	A	L	R	T	R	F	64	L	A	A	L	V	G	D	A	T	S	132
C₄	A	K	L	Q	G	G	M	I	F	H	15	R	H	P	A	L	R	T	S	F	61	F	H	H	A	V	I	D	G	W	S	144

			Motif-4							Motif-5												Motif-6					
	Y	X	D	F	A	V	W		V	G	X	F	V	N	T	L	Q	X	R		N	Q	D	Y	P	F	E	
				Y							I						Q	A				H						
C₁	Y	R	E	H	V	A	W	188	V	G	Y	Y	V	N	P	V	L	I	N	299	H	Q	A	Y	P	F	D	332
C₂	Y	R	D	F	V	A	Q	209	I	G	V	H	L	N	M	L	P	F	R	323	Y	R	Q	M	P	L	A	357
C₃'	Y	G	D	Y	A	C	W	173	I	G	F	F	I	N	Q	L	V	L	R	289	H	Q	H	V	P	F	D	322
C₃	Y	I	D	A	A	E	W	167	Y	A	E	L	A	N	A	Q	G	L	F	273	Q	Q	L	A	D	L	A	306
C₄	Y	H	E	F	V	R	L	181	L	G	L	F	I	N	S	I	P	L	R	292	H	R	R	Y	P	T	A	326

(B) PCP domains and ACP domain

	D		F	F	X	X	G	X	X	S		
PCP₁	D	D	F	F	A	L	G	G	H	S	L	30
PCP₂	D	N	F	F	D	L	G	G	D	S	I	30
PCP₃	D	N	F	F	A	L	G	G	N	S	I	30
PCP₄	D	N	F	F	A	L	G	G	D	S	I	30
ACP_L	L	P	I	Q	Q	L	G	L	D	S	L	30

Figure 4-45: NRPS gene locus-4 C and PCP domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS C domains (highlighted in yellow) from NRPS gene locus-4. The six core motif sequences as reported in the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules. The inactive domains are highlighted in grey.

(B) CLUSTALW alignment of the conserved core motif of NRPS PCP domains (highlighted in yellow) from NRPS gene locus-4. The core motif sequence as reported in the literature (Konz and Marahiel, 1999) is highlighted in bold. The PPant attaching serine residue is highlighted in green.

(A)

Motif-1 Motif-2 Motif-3 Motif-4 Motif-5

L S Y X E L   L K A G X A Y L P I D   L A Y X X Y T S G T T G X P K G   F D X S   N X Y G P T E

V   S

A₁ | L T F A D L | 31 | L K A G A T Y L P L D | 80 | P A Y I I Y T S G S T G Q P K G | 156 | F D L S | 193 | N L Y G P T E | 330
A₂ | I S Y Q Q L | 57 | L K A G A A Y L P L D | 106 | L A Y I I Y T S G S T G Q P K G | 181 | F D F S | 218 | N M Y G I T E | 319
A₃ | W S Y F D L | 62 | L Q L G A V Y L P I D | 111 | L A Y V L F T S G S T G T P K G | 181 | F D I S | 218 | N A Y G P A E | 319
A₄ | Y S Y Q A L | 49 | F Y A G A A Y L P L D | 98 | L A Y T L F T S G S T G T P K G | 181 | F D I S | 218 | N A Y G P A E | 319

Motif-6 Motif-7 Motif-8 Motif-9 Motif-10

G E L X I X G X G L A R G Y L   Y R T G D L   G R X D X Q V K I R G X R I E L G E I E   L P X Y M V P   N G K V D R

K   I   L

A₁ | G E L F I A G L G V A Q G Y L | 343 | Y K T G D L | 372 | G R N D Q Q V K I R G F R I E L G E I E | 405 | V P S Y M L P | 457 | N G K L D R | 477
A₂ | G E I Y V G G A G L A R G Y L | 382 | Y R S G D L | 410 | G R S D Q Q V K L R G F R I E L G E L E | 443 | L P H Y L Q P | 495 | N G K R D Y | 515
A₃ | G E L Y I A G I A V G W G Y L | 382 | Y R T G D L | 410 | G R S D Q Q V K I R G F R I E L G E I E | 443 | L P H Y L Q P | 495 | N G K L D R | 515
A₄ | G E L Y I G G V G V G R G Y L | 382 | Y R T G D L | 410 | G R A D Q Q V K V R G Y R I E L G E I E | 443 | L P S Y M L P | 492 | N G K L D R | 512

(B)

Gene locus	A domain	Stachelhaus code	Predicted substrates
haur_2108	A₁	DFWNIGMVH	Threonine
haur_2108	A₂	DILQLGLIWK	Glycine
haur_2106	A₃	DLTKVGEVGK	Asparagine
haur_2106	A₄	DSALVAEVW	Unknown

Figure 4-46: NRPS gene locus-4 A domains.

(A) CLUSTALW alignment of the conserved core motifs of NRPS A domains (highlighted in yellow) from NRPS gene locus-4. The ten core motif sequences as reported in the literature (Konz and Marahiel,1999) are highlighted in bold. The numbers represent the amino acid positions in the respective sequences. Suffices indicate the respective modules.

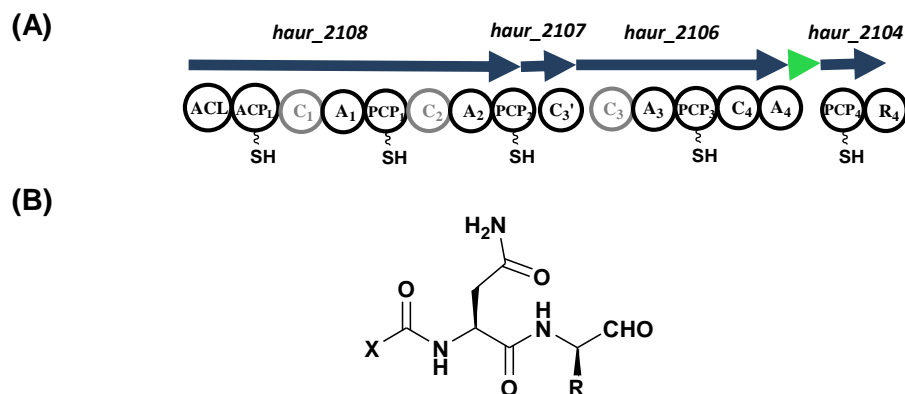(B) The table lists the Stachelhaus codes extracted from the sequences of NRPS gene locus-4 A domains and their putative substrates.

(A)

(B)

Figure 4-47: NRPS gene locus-4.

(A) Putative NRPS (blue) genes in NRPS gene locus-4. Green, accessory gene. Inactive domains are highlighted in grey.

(B)Predicted strcture of the putative dipeptide encoded by NRPS gene locus-4. X, Unknown fatty acid side chain; R, Unknown amino acid.

4.1.4.1.10 NRPS gene locus-5

The NRPS gene locus-5 with 9.5 Kb is rather small with only two NRPS modules. The conserved motifs required for the activity of the NRPS domains were identified by aligning the sequences (Figure 4-49). The A_L domain was predicted to activate tyrosine, whereas the A_1 domain may to activate asparagine resulting in a dipeptide (Figure 4-48). Sequences of the TE domain from module-1 lacks the catalytic triad S-89/D-116/H-264 required for its function (Figure 4-24). Hence the TE domain of the NRPS gene locus-4 is proposed to be inactive. The underlying mechanism in releasing the dipeptide remains unclear. In between the NRPS modules is the gene *haur_3128* encoding a glycosyltransferase for glycosylating the synthesized dipeptide (Table 4-13).

Downstream to the NRPS genes is *haur_3130* encoding for PPTase, which is necessary for priming the biosynthetic enzymes.

Gene locus	Modules	Putative functions
haur_3127	Module-L	A_L-PCP_L
haur_3128		Glycosyltransferases
haur_3129	Module-1	C_1-A_1-PCP_1-TE
haur_3130		Phosphopantetheinyltransferase

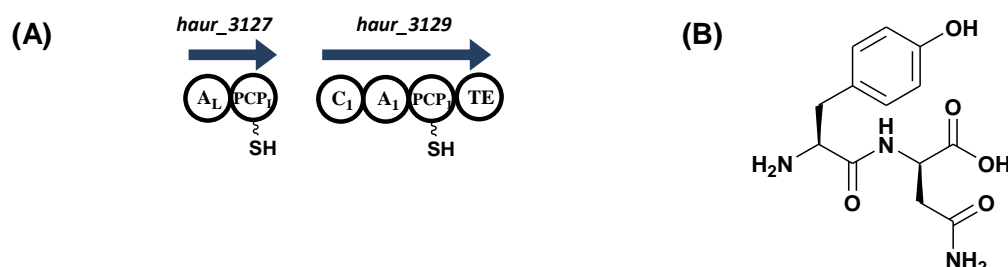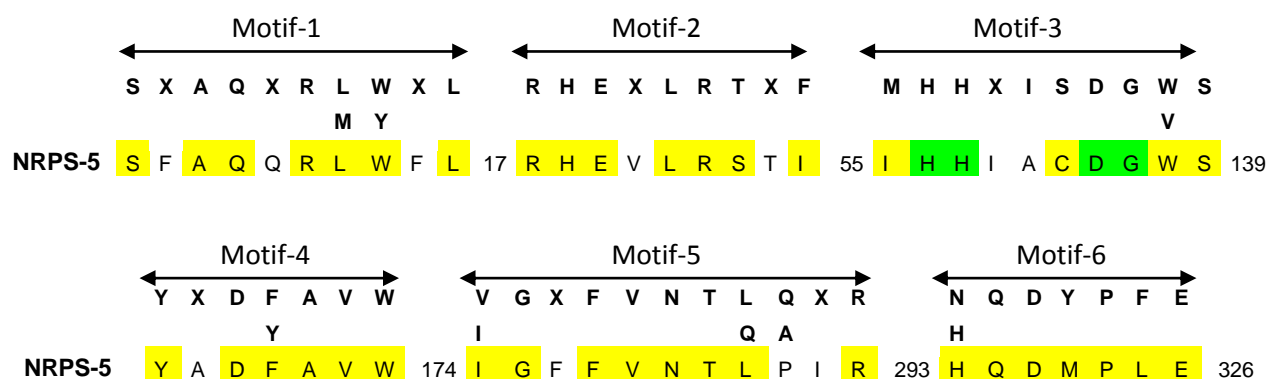**Table 4-13:** List of genes in NRPS gene locus-5. L, Loading module; A, Adenylation; PCP, Peptidyl Carrier Protein; C, Condensation; TE, Thioesterase.

Figure 4-48: NRPS gene locus-5.

(A) Putative NRPS (blue) genes in NRPS gene locus-5.

(B)Predicted strcture of the putative dipeptide encoded by NRPS gene locus-5.

(A) C domain

(B) PCP domains

(C) A domains

	Motif-1	Motif-2	Motif-3
	L S Y X E L	L K A G X A Y L P I D	L A Y X X Y T S G T T G X P K G
		V	S

A_L	L S Y A E F	26	W K A G A V Y L P L D	76	L A Y I I	Y T S G S T G K P K G	154
A_1	L S Y A E L	51	L K A G A A Y V P I D	101	L M Y V L	F T S G S T G R P K G	181

	Motif-4	Motif-5	Motif-6	Motif-7
	F D X S	N X Y G P T E	G E L X I X G X G L A R G Y L	Y R T G D L
			K	

A_L	F D A S	193	N A Y G P T E	282	G E L L V S S V G L A H G Y L	340	Y R T G D L	370
A_1	A D L G	220	N H Y G P T E	316	G E L Y V G G A G V G R G Y I	378	Y R S G D L	408

	Motif-8	Motif-9	Motif-10
	G R X D X Q V K I R G X R I E L G E I E	L P X Y M V P	N G K V D R
		I	L

A_L	G R I D H Q I K H R G Y R I D P G E I E	401	L P K Y M V P	456	N G K L D L	476
A_1	G R I D L Q V K I R G Y R V E L S E I E	439	L P D Y M V P	494	N G K I E R	514

Figure 4-49: Characteristic motifs in C (A), PCP (B) and A (C) domains of NRPS gene locus-5 (highlighted in yellow). The catalytically active residues are highlighted in green. The conserved residues from the literature (Konz and Marahiel, 1999) are highlighted in bold. The numbers indicate the position of amino acids. Suffices indicate the respective modules.

4.1.4.1.11 PKS gene locus

With the predominant occurrence of NRPS and PKS/NRPS mixed gene loci, only one PKS gene locus was identified from the *H. aurantiacus* DSM 785 genome sequence (Table 4-3). This PKS gene locus comprises KS, AT and KR domains but an ACP domain required for the binding of the substrate is missing. Albeit the sequence alignment of KS, AT, and KR domains showed the conserved motifs and catalytically active residues (Figure 4-20), a missing ACP domain (Table 4-14) is contemplated to render this PKS gene locus inactive.

Gene locus	Putative functions
haur_0866	KS-AT-KR

Table 4-14: PKS gene locus of *H. aurantiacus* DSM 785

4.2 Chemical screening of *H. aurantiacus* DSM 785

The computational analysis of the *H. aurantiacus* DSM 785 genome sequence gave an insight into the biosynthetic potential of this strain (section-4.1.4). Since siphonazole is the only known metabolite from a bacterium of the genus *Herpetosiphon* (Figure 4-50), isolation and structural elucidation of the products of the above described cryptic biosynthetic pathways (Table 4-3) would be of great interest.
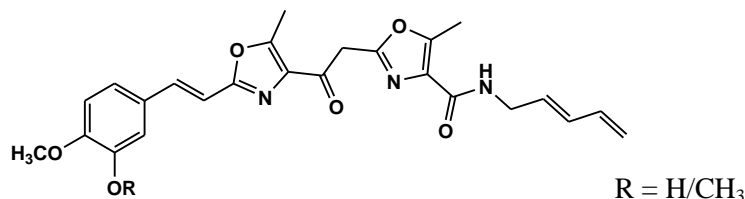

$$R = H/CH_3$$

Figure 4-50: Structure of siphonazole derivatives isolated from *Herpetosiphon* sp.

To investigate this, *H. aurantiacus* DSM 785 was cultivated in CY, MD1+G and HP74 medium as described in the section-3.2.4 (Figure 4-51). The culture was then centrifuged, and both the cell pellet and supernatant were analyzed further for the presence of secondary metabolites. The supernatant was extracted ethyl acetate, and this extract was fractionated by NP-VLC using dicholormethane, ethyl acetate, acetone and methanol as eluents. In parallel, the cell mass was extracted with methanol and fractionated in NP-VLC using dicholormethane, ethyl acetate, acetone and methanol. LC-MS and NMR analysis of the extracts showed molecular masses and NMR spectral features corresponding to the diketopiperazines cyclo(L-Pro-L-Tyr), cyclo(L-Pro-L-Val), cyclo(L-Hyp-L-Leu) and the indole derivative indol-3-yl-acetaldoxime (Figure-4.52). These diketopiperazines and the indole derivative were already reported to be by-products contained in the culture medium (Mitscher et al., 1967). Besides these compounds no trace of secondary metabolites could be identified from LC-MS and NMR analysis.
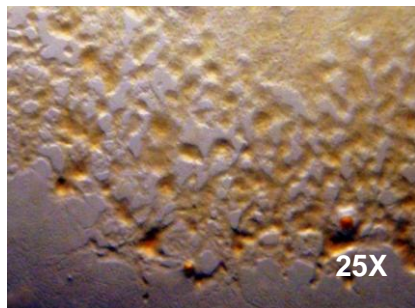
(A)   **(B)**

Figure 4-51: Swarm of strain *H. aurantiacus* DSM 785 on VY/2 agar (A) and the bacterium in CY medium (B).

Cyclo(L-Pro-L-Tyr)   Cyclo(L-Pro-L-Val)   Cyclo(L-Hyp-L-Leu)

Indol-3-yl-acetaldoxime
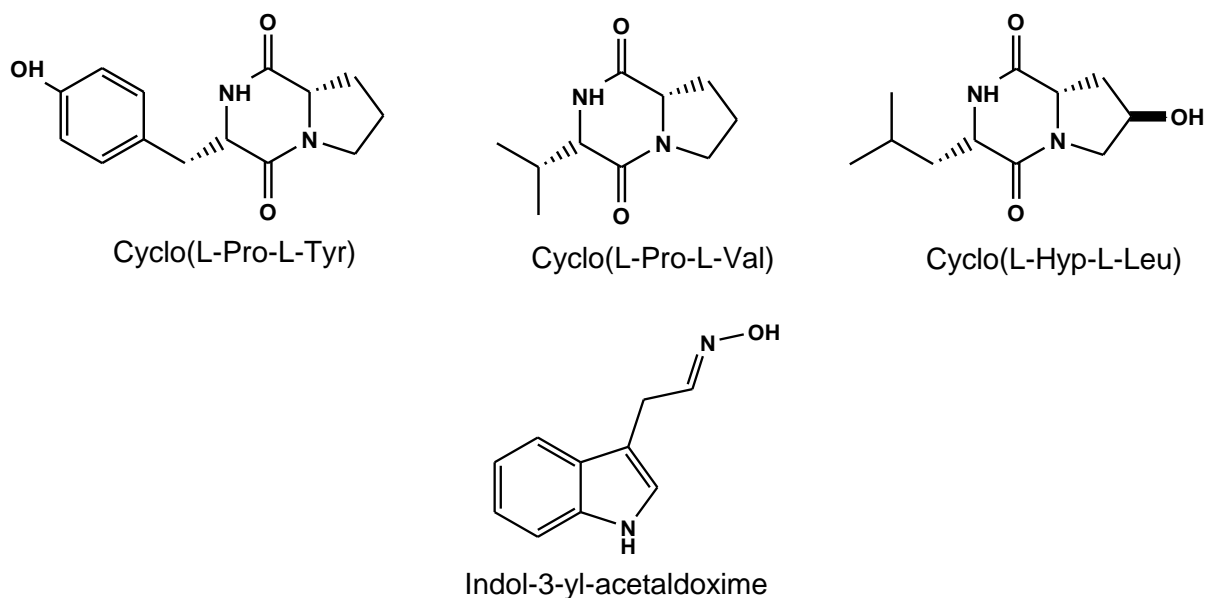
Figure 4-52: Structures of diketopiperazines and an indole derivative identified from *H. aurantiacus* DSM 785 culture extracts.

4.3 Exploring the genome based biosynthetic potential of *H. aurantiacus* DSM 785

The initial chemical screening of *H. aurantiacus* DSM 785 (section-4.2) did not revealed the products of the cryptic biosynthetic pathways analysed by our genome mining studies (Table 4-3). Thus, as a next step the expression of the cryptic biosynthetic genes was investigated under defined cultivation conditions.

4.3.1 Expression analysis for the PKS/NRPS gene locus-1

In the first step towards the identification of the putative lipopeptide encoded by PKS/NRPS gene locus-1 (section-4.1.4.1), the expression of PKS and NRPS genes was analysed at the transcriptional level. Total RNA was isolated from *H. aurantiacus* DSM 785, cultivated in CY and additionally in MD1+G medium for seven days. The isolated RNA was reverse transcribed using the forward primers of the primer pairs P_1, P_2 and P_3 designed from the PKS and NRPS sequences of the putative lipopeptide biosynthetic gene locus (Figure 4-53A). Since the PKS and NRPS genes (*haur_1858, haur_1859, haur_1860 and haur_1861*) are overlapping, the primer pairs, P_2 and P_3 were designed in such a way, that they span between the adjacent ORFs. The primer pair P_1 was from the PKS/NRPS hybrid module, *haur_1857* (Figure 4-53A). The synthesized cDNA was then PCR amplified by the respective primer pairs P_1, P_2 and P_3. Agarose gel electrophoresis of the amplified PCR product showed bands of the appropriate size only for the cDNA synthesized from the RNA isolated from *H. aurantiacus* DSM 785 cultivated in CY medium (Figure 4-53B). Sequencing of the amplicons confirmed the expression of the PKS/NRPS gene locus-1 in CY medium.

(A)

(B)
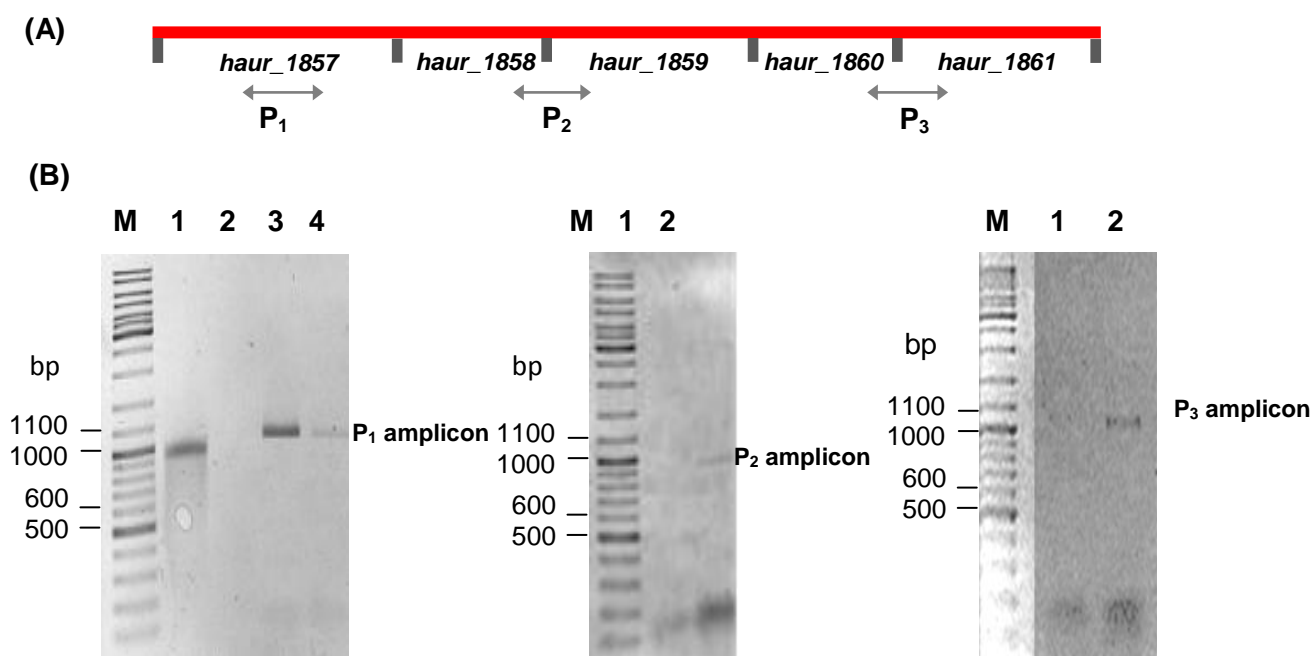
Figure 4-53: RT-PCR of PKS/NRPS gene locus-1.

(A) RT-PCR primer pairs, P_1, P_2 and P_3 designed specific for PKS and NRPS regions of PKS/NRPS gene locus-1. Red, PKS and NRPS genes in PKS/NRPS gene locus-1; black, ORFs. (B) Agarose gel electrophoresis of RT-PCR product amplified with the respective gene specific primers. Lane M, marker; Lane 1, RT-PCR positive control; Lane 2, 5 and 7,PCR amplified RNA using primer pairs P_1, P_2 and P_3 respectively; Lane 3, RT-PCR product using primer pair P_1 RNA (3µl); Lane 4, RT-PCR product using primer pair P_1, RNA (7µl); Lane 6, RT-PCR product using primer pair P_2, RNA (3µl); Lane 8, RT-PCR product using primer pair P_3, RNA (3µl).

4.3.2 Expression analysis of PKS/NRPS gene locus-2

The PKS/NRPS gene locus-2 with a PKS and eleven NRPS modules includes the genes encoding for the biosynthesis of HPG (Table 4-5). As mentioned in the section 4.1.4.2, one of the A domains, i.e A_2 was predicted to activate HPG (Figure 4-22B). Hubbard et al., 2012 described the biosynthesis of HPG by 4-hydroxyphenylpyruvate dioxygenase, p-hydroxymandelate synthase and aminotransferase. Kastner et al., 2012 reported the respective candidate genes *haur_1888, haur_1871, haur_1887* in *H. aurantiacus* DSM 785 responsible for the biosynthesis of HPG (Figure 4-21). Therefore, the expression of these genes was analyzed in *H. aurantiacus* 785 cultivated in CY and MD1+G medium for three, five and seven days. Since *haur_1887* and *haur_1888* were overlapping, the primer pair P_4 was designed for both ORFs. The primer pair P_5 was designed for the gene *haur_1871* (Figure 4-54A). The isolated RNA from the bacterium grown in CY medium was reverse transcribed and PCR amplified using the respective primer pairs. Agarose gel electrophoresis of the PCR amplified cDNA using the primer pair, P_5 showed a band of appropriate size from the RNA isolated after 2, 3 and 5 days (Figure 4-54B). But no band was observed from the RNA isolated after 7 days. Sequencing of the amplicons showed that *haur_1887* encoding for p-hydroxymandelate transaminase and *haur_1888* for p-hydroxyphenylpyruvate dioxygenase were expressing till day 5 but not any more days. cDNA synthesized from the primer designed from h*aur_1871* didn't show any PCR product in the agarose gel (Figure 4-54B). This confirmed that *haur_1871* encoding for p-hydroxymandelate synthase is not expressing in CY medium. This is the key enzyme in converting p-hydroxyphenylpyruvate to p-hydroxymandelate

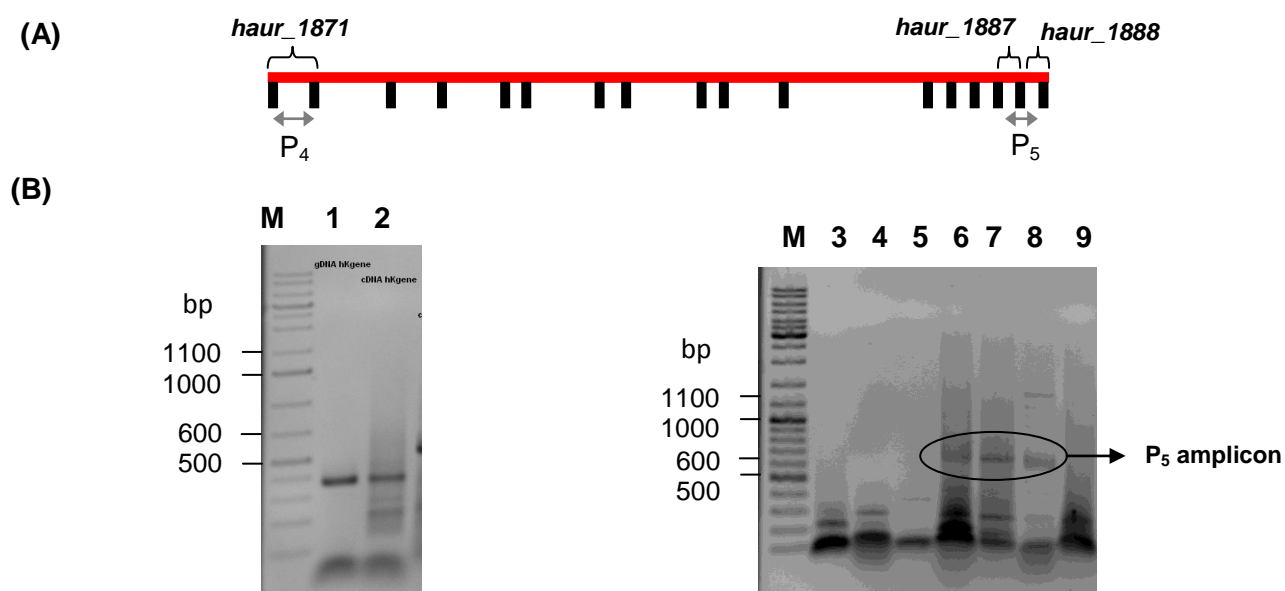(Figure 4-55). These results demonstrated that HPG biosynthesis is seized because *haur_1871* is not expressed.

(A)

(B)

Figure 4-54: RT-PCR of PKS/NRPS gene locus-2.

(A) RT-PCR primer pairs, P_4 and P_5 designed specific for HPG biosynthetic genes within PKS/NRPS gene locus-2. Red- PKS/NRPS gene locus-2; black, ORFs.

(B) Agarose gel electrophoresis of RT-PCR product amplified with the respective gene specific primers. Lane M-marker; Lane 1, PCR of *H. aurantiacus* DSM 785 genomic DNA using house keeping gene specific primers P_6 (*haur_0866*); Lane 2, RT-PCR product using house keeping gene specific primers P_6 (positive control); Lane 3, 4 and 5, RT-PCR product using primer pair P_4 and RNA isolated after 2, 5 and 7 days, respectively; Lane 6, 7, 8 and 9, RT-PCR product using primer pair P_5 and RNA isolated after 2, 3, 5 and 7 days, respectively.

126
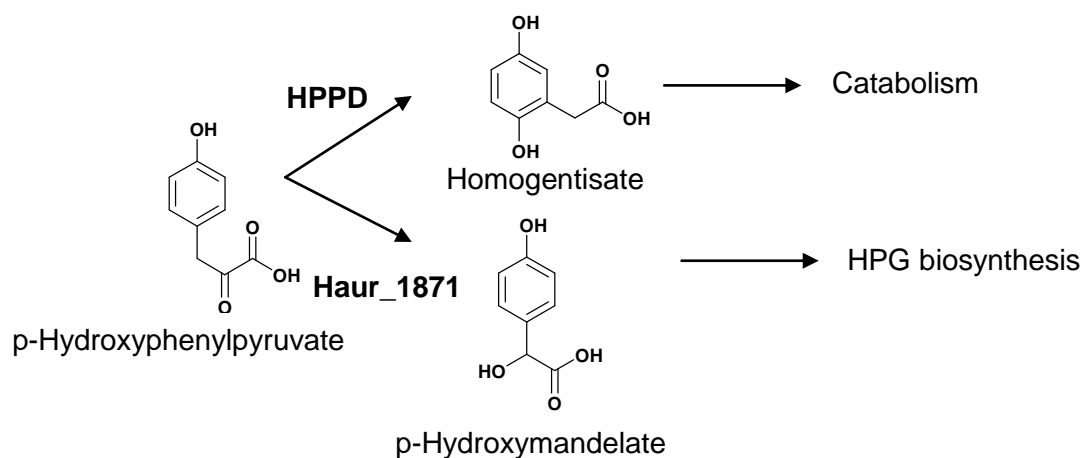
Figure 4-55: Metabolism of p-hydroxyphenylpyruvate. HPPD, p-hydroxyphenylpyruvate dioxygenase; HPG, p-hydroxyphenylglycine.

4.3.3 Expression analysis of NRPS gene locus-3

NRPS gene locus-3 with five NRPS modules is organised in five ORFs (Table 4-11). The locus *haur_2089*, which is a part of the NRPS gene locus-3, carries an NRPS module with a TE domain for releasing the synthesized peptide from the enzyme assembly (Table 4-11). The primer pair P_7 was designed for this termination module encoding region (Figure 4-56A). In addition, another primer pair, P_8 was designed for *haur_2091* (Figure 4-56A), which is considered to encode an enzyme which performs the second elongation step in the biosynthesis of the putative peptide. The RNA isolated from *H. aurantiacus* DSM 785 cultured in CY and MD1+G medium for 7 days was reverse transcribed using primer pairs P_7 and P_8. PCR of the synthesized cDNAs using both the primer pairs did not show any bands (Figure 4-56B). The inability of amplifying cDNAs using the respective primers showed that the NRPS gene locus-3 is not expressing in CY medium.

4.3.4 Expression analysis of NRPS gene locus-4

NRPS gene locus-4 with NRPS genes organised in four ORFs was described in section-4.1.4.9, to encode for a putative lipodipeptide (Figure 4-47). For the expression analysis a single primer pair P_9 was designed for the intiation module encoding gene, *haur_2108*. cDNA was synthesized from the RNA isolated from *H. aurantiacus* DSM 785 cultivated in CY and MD1+G medium. PCR of the reverse transcribed product was done using the designed gene specific primer pairs P_9. Agarose gel electrophoresis of the PCR product showed band of the expected size of 730 bp in the cDNA synthesized from the RNA isolated from *H. aurantiacus* DSM 785 in CY medium (Figure 4-56B). Sequencing of the amplicons confirmed the expression of the gene in CY medium. Therefore, it is expected that the NRPS gene locus-4 expresses in CY medium.
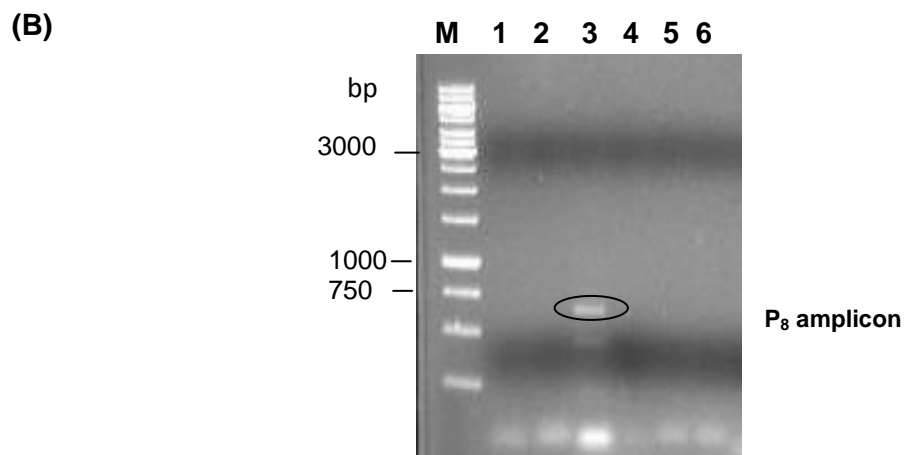
(A)

(B)

Figure 4-56: RT-PCR of NRPS gene locus-4.

(A) RT-PCR primer pairs, P_7 and P_8 designed specific for NRPS gene, *haur_2089* and *haur_2091*. Red, NRPS genes of NRPS gene locus-3; black, ORFs.

(B) Agarose gel electrophoresis of RT-PCR product amplified with the respective gene specific primers. Lane M, marker (Promega, 1 Kb DNA ladder); Lane 1, primer pair of NRPS gene locus-4 and RNA from *H. aurantiacus* DSM 785 in MD1+G medium; Lane 2- 3, primer pairs P_7 and P_8 of NRPS gene locus-3 and RNA from *H. aurantiacus* DSM 785 in MD1+G medium; Lane 4- primer pair P_9 of NRPS gene locus-4 and RNA from *H. aurantiacus* DSM 785 in CY medium; Lane 5-6, primer pairs P_7 and P_8 of NRPS gene locus-3 and RNA from *H. aurantiacus* DSM 785 in CY medium.

4.3.5 Strategies to uncover *H. aurantiacus* DSM 785 secondary metabolites, focussing on the product of PKS/NRPS gene locus-1

Examining the expression of cryptic biosynthetic genes, PKS/NRPS gene loci-1 and 2, NRPS gene loci 3 and 4, showed that CY medium is the preferred culture medium for the expression of PKS/NRPS gene locus-1 and NRPS gene locus 4. Therefore, *H. aurantiacus* was cultivated in CY medium as described in section-2.1. However, the preliminary chemical screening of such *H. aurantiacus* culture had not resulted in any secondary metabolites (Section-3.2). Therefore, different extraction methodologies were adopted. Also, the CY medium was supplemented with additives in order to probably enhance the biosynthesis of secondary metabolites in *H. aurantiacus*.

The focus was put on PKS/NRPS gene locus-1, which was described in section- to encode for a putative lipopeptide (Figure 4-18). It was regarded as an advantage, that a lipopeptide is expected to posses' anti-microbial activity. The PKS and NRPS genes of PKS/NRPS gene locus-1 could be analysed with some certainty concerning the modules and their function (Section-4.1.4.1).

4.3.5.1 Extraction methods

There were reports (Akhihiro et al., 1991; Liu et al., 2008) claiming the isolation of lipopeptides from both, the broth and cell mass. Hence, the supernatant and cell pellet obtained from the *H. aurantiacus* cultivated in CY medium were extracted further for the isolation of secondary metabolites (Figure 4-57). The extraction procedures were guided by LC-MS analysis of the extracts.

H. aurantiacus DSM 785 in CY medium for 7 days

Supernatant

Cell mass
(Liu et al., 2008)

Aqueous phase   Organic phase
(Ohno et al., 1995; Borders et al., 2004)
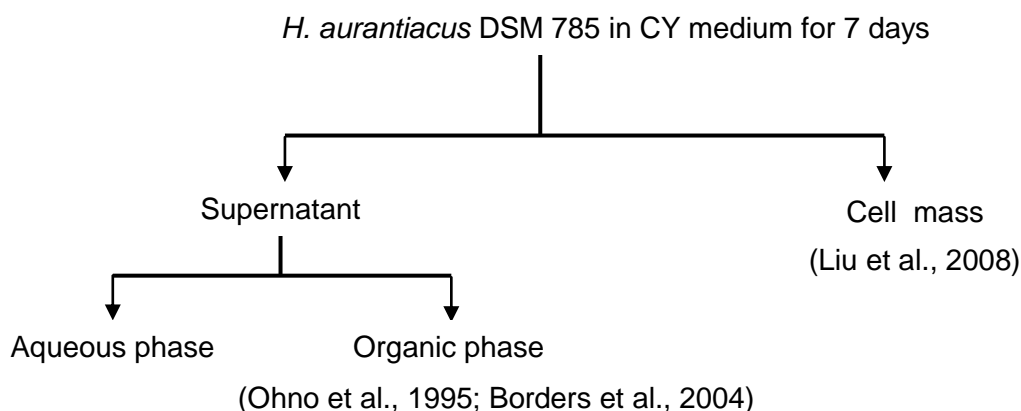
Figure 4-57: Extraction scheme for *H. aurantiacus* culture

Supernatant extract and its chemical analysis

First, the *H. aurantiacus* culture supernatant was extracted with butanol. The butanol extract was mixed with water and adjusted to pH-2 using HCl. This suspension was then allowed to separate and the organic phase was collected. The collected organic phase was washed with water to get rid of polar impurities. The washed organic phase was further fractionated using NP-VLC using dichloromethane, dichloromethane:acetone (1:1), acetone, acetone:methanol (1:1) and methanol as eluent to obtain five fractions.

In addition to the above method, the butanol extract was evaporated under vacuum and subsequently are fractionated by NP-VLC using dichloromethane, dichloromethane:acetone (1:1), acetone, acetone:methanol (1:1) and methanol as eluents.

In a further experiment, precipitation of the lipopeptides was attempted. As in the isolation of surfactin (Ohno et al., 1995), the *H. aurantiacus* culture supernatant was adjusted to pH-2 using HCl. The mixture was incubated overnight at 4°C to allow any lipopeptides to precipitate. The precipitate was then collected by centrifuging. Precipitated material was in very low amount. It was extracted with methanol and evaporated under vacuum.

LC-MS and NMR analysis of all above mentioned extracts and fractions showed molecular masses and NMR spectral features corresponding to diketopiperazines and the indole derivative described in section-4.2.

Cell mass extract and its chemical analysis

The cell pellet obtained from *H. aurantiacus* in CY medium (4 L) was lyophilized and suspended in a methanol and dichloromethane (2:1) mixture. The suspension was then centrifuged and the supernatant collected. The collected supernatant was evaporated under vacuum. This crude methanol extract was fractionated by NP-VLC using dichloromethane, ethyl acetate, acetone and methanol as eluents. LC-MS analysis of this extract showed an interesting molecular mass of approximately 774 Da (Compound-1). HR-MS of this crude methanol extract using a RP C-18 column and acetonitrile/water/formic acid as mobile phase showed that the exact mass of compound-1 as 772.9128 Da (Figure-8.3). Therefore, the crude methanol extract was further purified in RP-18 so as to isolate the putative compound-1. 12 fractions were collected using gradient elution starting from 50:50 (methanol:water) to 100% methanol. LC-MS analysis of the 12 fractions showed the presence of compound-1 in the fraction-8. Since fraction-8 comprised only 3 mg and still was a mixture, the cultivation was scaled up to 40 L.

The resulted crude methanol was fractionated by NP-VLC as described above. RP-HPLC of the NP-VLC methanol fraction using C8 column with acetonitrile/water (7:3) resulted in 6 fractions. LC-MS analysis of the RP-HPLC fractions showed that the compound-1 was distributed throughout all the fractions. In addition to compound-1 with a mass of 774 Da, another compound-2 with a molecular mass approximately 745 Da was also identified in all the fractions(Figure-8.2). Therefore, all RP-HPLC fractions composed of the molecular masses 774 Da and 745 Da were pooled again and it was decided to first to get rid of impurities that prevent the purification of compounds-1 and 2.

To get rid of any polar impurities, the extract was fractionated further using RP-VLC and different water and methanol mixture (9:1, 3:1, 1:1, 1:3, 0:1) as eluents. LC-MS analysis of the fractions showed that the RP-VLC fraction obtained with 100% methanol contained compounds-1 and -2. The methanol fraction was separated further using size exclusion chromatography on sephadex LH-20. 14 fractions were collected and analyzed by LC-MS, and fractions 6-9 were identified to include the compounds-1 and -2.

NMR analysis of the pooled fractions obtained from size exclusion chromatography showed that they contain large amounts of lipids. Therefore, the extract was liquid-liquid partitioned between chloroform and methanol:water (7:3) to remove lipids. But, LC-MS analysis of the both phases showed that the chloroform phase contained compounds-1 and -2. RP-HPLC of the chloroform extract was not successful in purifying compounds-1 and -2. Nevertheless, the extract was fractionated in SPE columns by gradient elution starting from dichloromethane and methanol (9:1, 8:2, 7:3 6:4, 5:5, 4:6, 3:7, 2:8, 1:9) to 100% methanol. LC-MS analysis of the fractions showed that the compounds 1 and -2 co-eluted in fraction 4 with (6:4) dichloromethane and methanol (Figure-8.2). The amount of the fraction-4 obtained from the above stages of purification was 1 mg. The amount was too low to proceed with further purifications. Therefore, the structure of compounds-1 and -2 could not be elucidated in detail (Figure-8.1).

Compounds-1 and -2 have a characteristic UV absorption maximum at 376 nm and 354 nm respectively. In the LC-MS chromatogram, compounds-1 and -2 were found to co-elute with UV absorption at 362 nm (Figure-8.2).

4.3.5.2 Modification of cultivation methods to facilitate the biosynthesis of secondary metabolites by *H. aurantiacus*

It was decided to change the cultivation media composition in order to increase the amounts of compounds-1 and -2 and possibly the isolation of the putative lipopeptide and other secondary metabolites.

Effect of HPG on the metabolic profile of H. aurantiacus DSM 785

As per the NRPS predictor tool, the A_3 domain of PKS/NRPS gene locus-1 is expected to activate HPG (Figure 4-8B). Moreover, A_2 of PKS/NRPS gene locus-2 and A_1 of NRPS gene locus-1 was predicted to activate HPG (Figure 4-22B). Expression of PKS/NRPS gene locus-1 in CY medium and the failure of HPG biosynthetic genes to express in CY medium were used to refine the cultivation strategies to isolate the putative lipopeptide.

In the first attempt, 3 days old *H. aurantiacus* DSM 785 culture cultivated in 3 L CY medium was supplemented with 1mM of tyrosine. Tyrosine, the precursor of HPG, is anticipated to induce the expression of HPG biosynthetic genes. Since, HPG is expected to be one of the building blocks of the putative lipopeptide the expression of the HPG biosynthetic genes is considered to increase the production of the putative lipopeptide as well. At the same time, *H. aurantiacus* DSM 785 was cultivated in 3 L of CY medium supplemented with 3mM of L-HPG. The supernatant and cell mass were extracted from the cultures using the procedures described in the section-4.4.1. LCMS and NMR spectra of the extracts were similar to that of the extracts isolated from CY medium without tyrosine and HPG. Thus, the supplementation of tyrosine or HPG into CY medium didn't influence the secondary metabolic profile of *H. aurantiacus* DSM 785.

Bacteriolysis and the metabolic profile of H. aurantiacus DSM 785

Lysis of dead *E. coli* cells by *H. aurantiacus* DSM 785 similar to myxobacteria was reported earlier (Trick and Lingens, 1984). Therefore, cultivation of *H. aurantiacus* DSM 785 in the presence of dead *E. coli* cells is speculated to bring some metabolic changes in *H. aurantiacus* DSM 785. This is in turn is expected to reflect in the metabolic profile of *H. aurantiacus* DSM 785 as well. Therefore, 10 ml of dead *E. coli* suspension prepared according to the section-3.2.4 was added to 1.5 L of 3 days old *H. aurantiacus* DSM 785 in CY medium. The supernatant and cell mass were extracted as described in the section-4.4.1. In contrary to expected, no difference in the metabolic profile was observed in *H. aurantiacus* DSM 785 cultivated along with dead *E. coli* suspension in CY medium.

From gene expression studies using RT-PCR and the growth behaviour, CY medium was opted as the better medium to cultivate *H. aurantiacus* DSM 785. To extract the biosynthetic potential of *H. aurantiacus* DSM 785, the strain was cultivated in CY medium along with additives including tyrosine, L-HPG and the dead *E. coli* suspension. Moreover, different extraction methods were followed for the above cultivation conditions. Albeit, two compounds-1 and 2 were identified in the extract neither of the compounds were in good amounts to proceed further. None of the other metabolites encoded by the PKS and/or NRPS genes of *H. aurantiacus* DSM 785 were also isolated.

4.4 Biochemical characterization of A domains in PKS/NRPS gene locus-1

PKS/NRPS gene locus-1 comprises six A domains responsible for selecting amino acids for the putative lipopeptide (Table 4-4). As mentioned in section-4.4.1.1, the putative substrates activated by all A domains except for the A_2 domain, were anticipated from their amino acid sequences (Figure 4-8B). In order to verify these deductions the A domains were biochemically characterized. Moreover, using this approach it was attempted to identify the substrate of the A_2 domain. For the characterization, overexpressed A domain proteins were assayed using γ-$^{18}O_4$- ATP pyrophosphate exchange assay in the presence of predicted and other substrates (Phelan et al., 2009).

4.4.1 Heterologous expression and purification of the A_3 domain

To initiate the biosynthesis of the full length N-terminal His_6-tagged A_3 domain protein, 2.3 Kb of the nucleotide sequence encoding for the A_3 domain was first PCR amplified using *pfu* DNA polymerase from the isolated *H. aurantiacus* DSM 785 genomic DNA. The amplicon with appropriate restriction sites was then cloned into the pET-28a expression vector. For the overexpression, the generated construct pLN-A_3 was transformed into the *E. coli* BL21 cells. The recombinant *E. coli* BL21-pLN-A_3 cells were then cultivated in LB medium with kanamycin till an OD_{600} of 0.5 was reached. The culture was then induced with different concentrations of IPTG to enable the expression of the His_6-tagged A_3 domain protein and incubated further at different temperatures for varying time period. However, the protein did not express in the above conditions tested.

Owing to the codon usage bias between *E. coli* and *H. aurantiacus* DSM 785, the plasmid pLN-A_3 was transformed into Rosetta (DE3) pLysS cells. Rosetta (DE3) pLysS facilitates the expression of genes that contain the codons rarely used in *E. coli*. The cultivation of Rosetta-pLNA$_3$ cells in LB medium with kanamycin till OD_{600} of 0.5 followed by the induction with different concentrations of IPTG and incubation at 37°C showed the expression of a 63.6 kDa His_6-tagged A_3 protein in the insoluble fraction (Figure 4-58A). The protein was then successfully expressed in the soluble fraction from the cells induced with 0.5 mM IPTG and post-induction incubation at 16°C for 16 hrs. Due to the low level of expression the protein was extracted from 1L LB medium for the adenylation assay.

The expressed protein was then purified via affinity chromatography using gravity flow Ni-NTA resin (Figure 4-58B). The purified protein fractions were then desalted and exchanged in Tris buffer pH-7.5 using PD-10 desalting column. The protein fractions were further concentrated using 10 kD cut-off Amicon centrifugal concentrator.
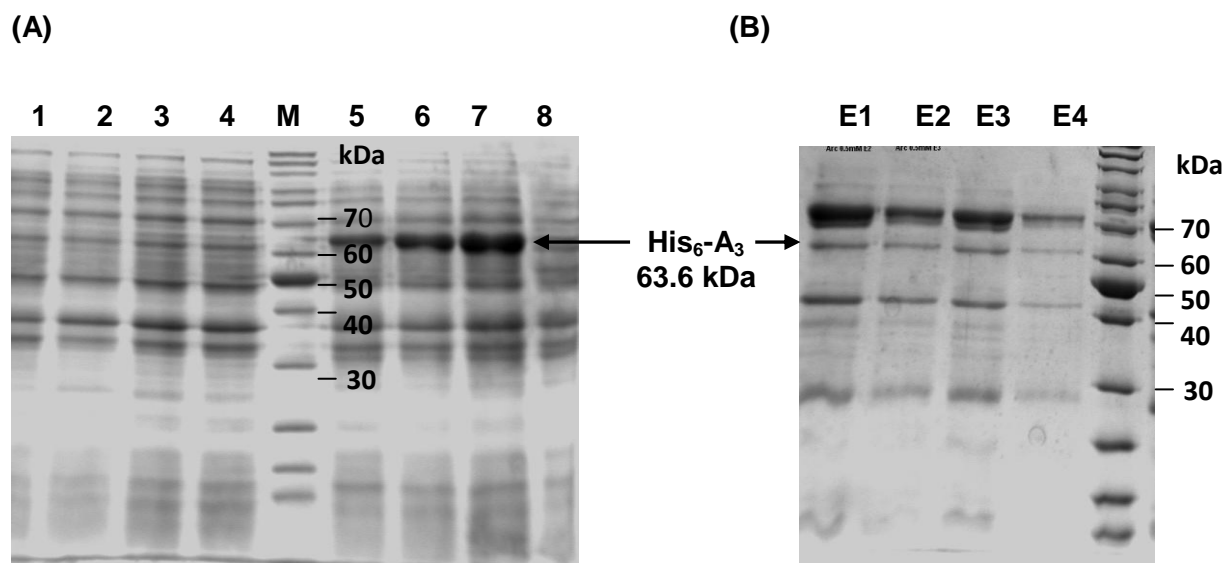
(A)   **(B)**

Figure 4-58: Heterologous expression and purification of His$_6$-A$_3$ protein.

(A) SDS-PAGE (11%) of protein extracts from Rosetta (DE3) pLysS-pLNA$_3$ at 37°C. Lane 1-Control (IF); Lane 2, 3 and 4- 1mM IPTG, 0.5mM IPTG and 0.25mM IPTG (IF); Lane 5, 6 and 7- 0.25 mM IPTG, 0.5mM IPTG and 1mM IPTG (SF); Lane 8-Control (SF). IF-Insoluble fraction, SF-Soluble fraction. M- marker.

(C) SDS-PAGE (11%) of Ni-NTA elution fractions (E1-E4) containing His$_6$-tagged A$_3$ protein.

4.4.1.1 Adenylation assay with A$_3$ domain

Bioinformatic analysis of A$_3$ protein sequences showed that A$_3$ domain of *haur_1858* should preferentially activate the amino acid, L-asparagine. To support the bioinformatic prediction, the heterologously expressed A$_3$ domain protein was analysed using γ -^{18}O$_4$-ATP pyrophosphate exchange assay in the presence of L-asparagine. In addition to L-asparagine other amino acids showing similar physiochemical properties like L-aspartate, L-glutamine and L-glutamate and other diverse amino acids like L-histidine and L-cysteine were also tested. MALDI-TOF MS of the assay mixture showed activation of all the tested amino acids without any substantial difference (Figure 4-59).

To improve the function of the A$_3$ domain protein, the assay was carried out at 30°C for 30 min. This was done because *H. aurantiacus* DSM 785 grows well at 30°C. Analysis of the assay mixtures showed that at this temperature the A$_3$ domain protein was inactive. Since the A$_3$ domain protein was eluted along with some background proteins,

the influence of these background proteins on unspecific activation of substrates by His-tagged A_3 domain protein was examined. Thus, the assay was done with proteins extracted from the uninduced culture. But none of the amino acids tested were activated. Therefore, it is most likely that indeed the A_3 domain protein is responsible for the unspecific activation.
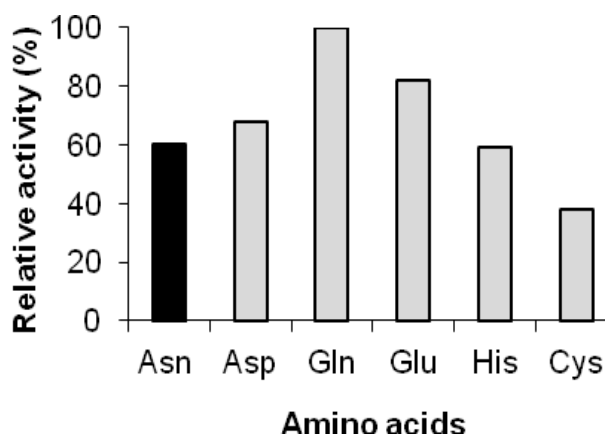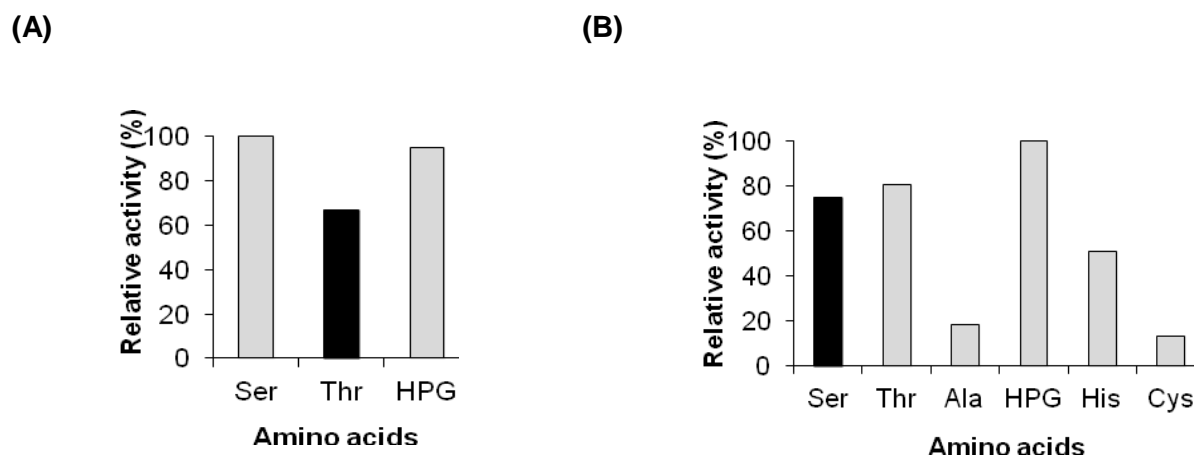

Figure 4-59: γ -$^{18}O_4$-ATP pyrophosphate exchange assay with His_6-tagged A_3 protein. Black bar indicates activity of A_3 domain with the predicted substrate (Asn). Grey bars indicate activity of A_3 domain with other substrates used in the assay.

4.4.2 Heterologous expression, purification of A_4 and A_5 domains

The bimodular locus *haur_1859*, which is a part of PKS/NRPS gene locus-1, has two A domains, i.e A_4 and A_5 (Table 4-3). In order to assess the activity of these A domains, the nucleotide sequences encoding for the A_4 and A_5 domains were cloned into the pET28 expression vector to generate the constructs $pLNA_4$ and $pLNA_5$. The expression of N-terminus His_6-tagged A_4 and A_5 domain proteins was achieved using the same procedure as that described for $pLNA_3$ with 0.5 mM IPTG and a post-induction incubation at 16°C for 18 hrs. The purification of the A_4 and A_5 domain proteins in their native form from the soluble fraction was carried out using gravity flow Ni-NTA resin. SDS-PAGE of the protein fractions followed by the Coomassie blue staining and destaining showed a band of 49.7 kDa and 48.7 kDa size in consistent with the predicted molecular weight of the His_6-tagged A_4 and A_5 proteins, respectively. The purified protein fractions were then desalted and concentrated (Figure 4-60).

(A)

(B)
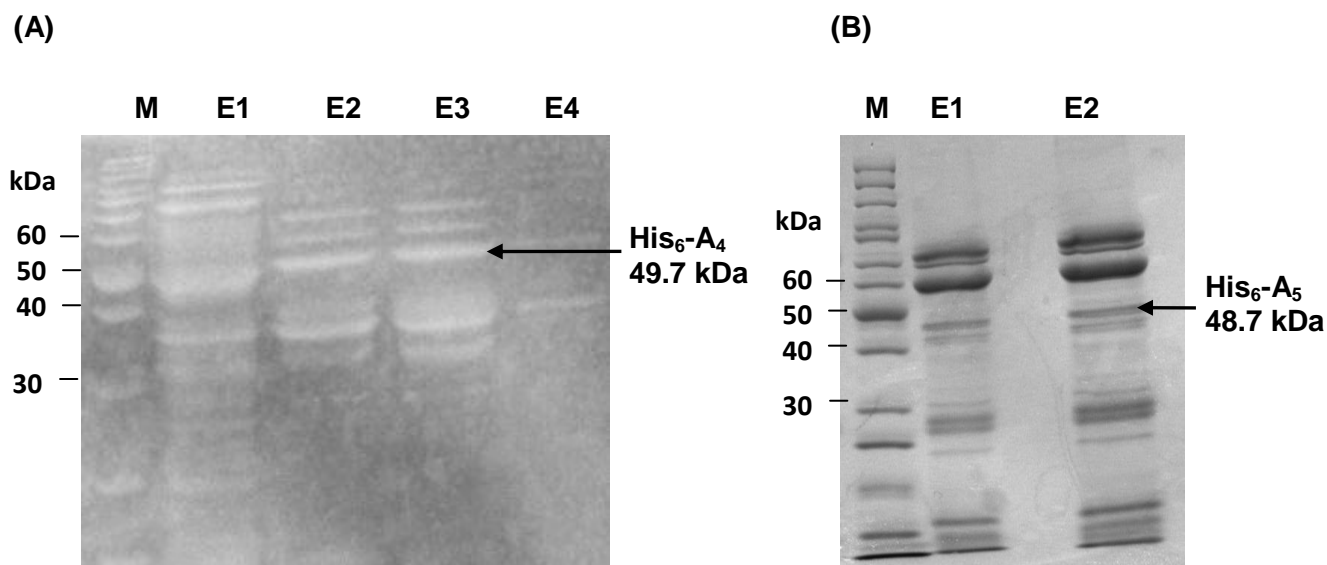
Figure 4-60: Heterologous expression and purification of His_6-A_4 and His_6-A_5 proteins.
(A) SDS-PAGE (12%) of Ni-NTA elution fractions (E1-E4) containing His_6-tagged A_4 protein. M-marker.
(B) SDS-PAGE (11%) of Ni-NTA elution fractions (E1-E2) containing His_6-tagged A_5 protein.

4.4.2.1 Adenylation assay with A_4 and A_5 domains

The Stachelhaus code for the A_4 and A_5 domains were congruent with that for L-threonine and L-serine, respectively (Figure 4-8B). Since serine and threonine differ only in a methylene group, the expressed A_4 and A_5 domain proteins were examined in the presence of L-serine and L-threonine using the γ $^{18}O_4$-ATP pyrophosphate exchange assay. In addition other amino acids including L-alanine, L-histidine, L-cysteine and L-HPG were also assayed. But again, as in the case of A_3 domain both the A_4 and A_5 domain proteins were highly promiscuous (Figure 4-61).

(A)   **(B)**

Figure 4-61: (A) γ-$^{18}O_4$-ATP pyrophosphate exchange assay with His$_6$-tagged A_4 protein. Black bar indicates activity of A_4 domain with the predicted substrate (Thr). Grey bars indicate activity of A_4 domain with other substrates used in the assay.

(B) γ-$^{18}O_4$-ATP pyrophosphate exchange assay with His$_6$-tagged A_5 protein. Black bar indicates activity of A_5 domain with the predicted substrate (Ser). Grey bars indicate activity of A_5 domain with other substrates used in the assay.

Although the tested A domains, i.e A_3, A_4 and A_5 were functional they were unspecific. These results demonstrated that discrete A domain proteins of PKS/NRPS gene locus-1 of *H. aurantiacus* DSM 785 were either indeed unspecific or a larger protein is required to gain specific activity.

4.4.3 Heterologous expression with A_3-PCP$_3$ didomain

Marshall et al., 2002 reported the significance of interdomain interactions between *trans* acting A and PCP domains in recognizing their substrates. Therefore, it is expected that the respective PCP domains might improve the in vitro behaviour of A_3, A_4 and A_5 domain proteins. To generate N-terminus His$_6$-tagged A_3-PCP$_3$ didomain protein, the encoding nucleotide sequences were cloned into the pET28a expression vector resulting in the construct pLNA$_3$-PCP$_3$. For the expression and purification of the A_3-PCP$_3$ didomain protein the same procedure as that for the discrete A_3 domain was followed, except that after induction with IPTG the culture was incubated at 18°C for 24 hrs. Coomassie blue staining and destaining of the SDS-PAGE from the eluted protein fractions showed a band of 75.9 kDa, corresponding with the expected molecular weight

of the N-terminus His$_6$-tagged A$_3$-PCP$_3$ didomain protein. The heterologously expressed protein was then purified using gravity flow Ni-NTA resin (Figure 4-62). The eluate was desalted and concentrated.
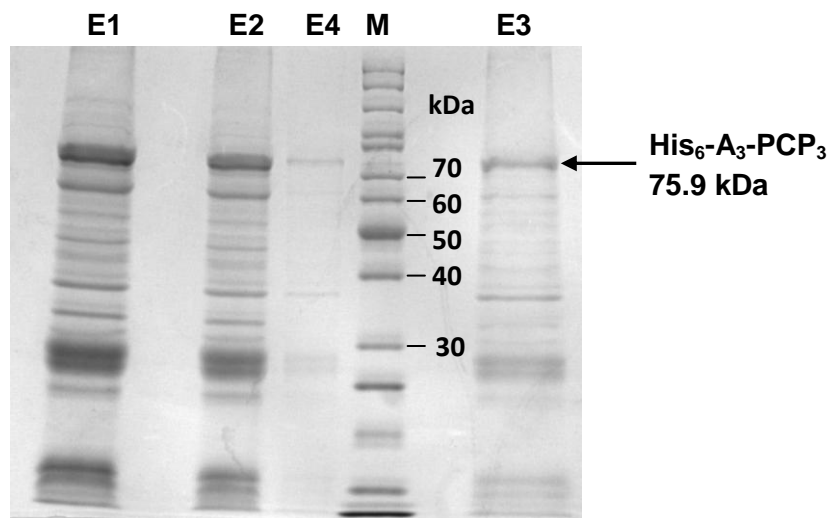

Figure 4-62: Heterologous expression and purification of His$_6$-A$_3$-PCP$_3$. SDS-PAGE (11%) of Ni-NTA elution fractions (E1-E4) containing His$_6$-tagged A$_3$-PCP$_3$ protein. M-Marker.

4.4.3.1 Adenylation assay with A$_3$-PCP$_3$ didomain

The γ-^{18}O$_4$-ATP pyrophosphate exchange assay of the concentrated A$_3$-PCP$_3$ didomain protein was done with the same set of amino acids as for the A$_3$ domain alone; L-asparagine, L-aspartate, L-glutamine, L-glutamate, L- histidine and L-cysteine. L-asparagine was highly activated but other amino amino acids were also equally activated (Figure 4-63). This showed that the PCP domain did not influence the activity of the A$_3$ domain.
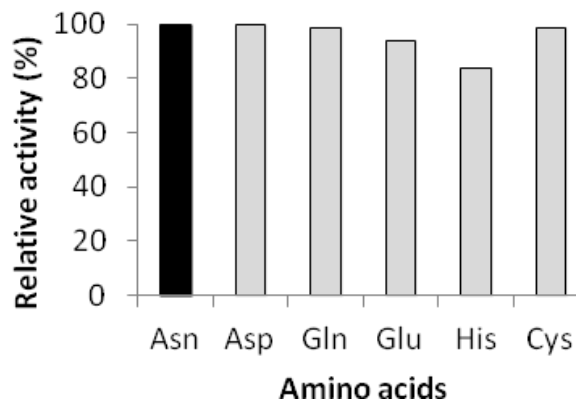
Figure 4-63: γ-$^{18}O_4$-ATP pyrophosphate exchange assay with His$_6$-tagged A$_3$-PCP$_3$ protein. Black bar indicates activity of A$_3$ domain with the predicted substrate (Asn). Grey bars indicate activity of A$_3$ domain with other substrates used in the assay.

4.4.3.2 Heterologous expression of the PPTase Haur_3130

In a further attempt a mass spectral characterization of the loaded amino acids was tried. For this a PPTase was needed. To identify the covalently loaded substrates, the expressed A-PCP didomains needed to be incubated with PPTase. PPTase is the primary enzyme for converting the PCP domain to active holo form. This in turn renders the PCP domain to bind the adenylated subtrates. The *H. aurantiacus* DSM 785 genome sequence encoding a PPTase at the locus *haur_3130* for priming the PCP domains of NRPS megasynthases.

To generate a full length N-terminal His-tagged PPTase, *haur_3130* was cloned in pET28a to generate the construct pLNPPT. The construct was then transformed in to the expression host BL21 (DE3) pLysS. 30 kDa of the His-tagged PPTase was then successfully expressed in of LB medium with 0.5mM IPTG and post-induction incubation at 16°C for 16 hrs (Figure 4-64). The N-terminal his-tagged PPTase was purified using gravity flow Ni-NTA resin. The purified protein was then desalted and concentrated further.
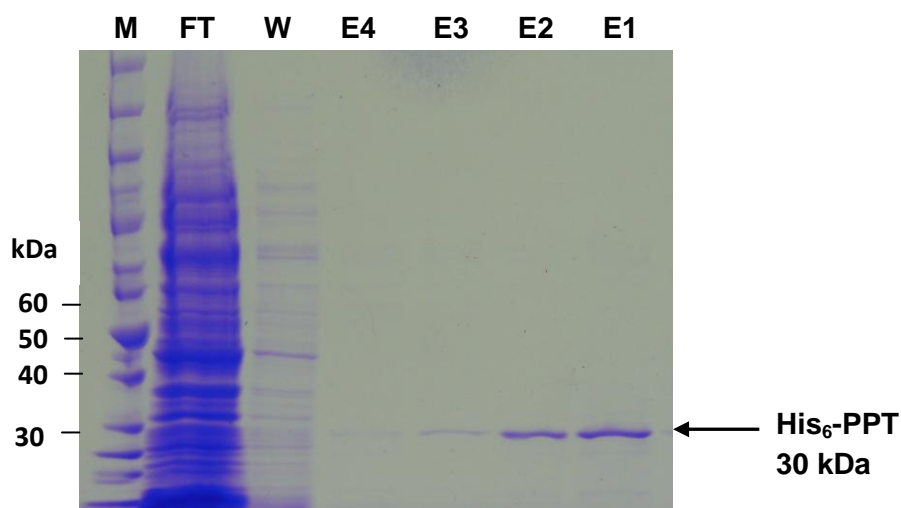
Figure 4-64: Heterologous expression and purification of His$_6$-tagged PPT. SDS-PAGE (11%) of Ni-NTA purified fractions containing His$_6$-tagged PPT. PPT, phosphopantetheinyltransferase; M, marker; FT, flow through; W, wash fraction; E1-E4, elution fractions.

4.4.3.3 Phosphopantetheinylation of A$_3$-PCP$_3$ didomain

Heterologously expressed A$_3$-PCP$_3$ didomain was incubated with Haur_3130 and amino acid substrates, in order to analyze the loading of specific substrates onto the PCP$_3$ domain. For the beginning, native SDS-PAGE of the assay mixtures including A$_3$-PCP$_3$(apo), A$_3$-PCP$_3$(holo) and A$_3$-PCP$_3$(holo) incubated with amino acid substrates showed bands at 75.9 kDa corresponding to the A$_3$-PCP$_3$ didomain and 30 KDa for Haur_3130. The mass difference between the A$_3$-PCP$_3$(apo) and A$_3$-PCP$_3$(holo) with and without the substrate loading is too small to be identified from the SDS-PAGE. The in-gel digestion of the 75.9 kDa band and subsequent analysis using MALDI TOF/TOF (Section-3.5.8.2) was expected to reveal the mass difference and protein sequence as well. The peptide fingerprints showed the presence of catalytically active serine for the loading of phosphopantetheine arm in one of the fragment of molecular weight 3346.82 Da. But, shift in the molecular mass to 3686.92 Da.corresponding to the A-PCP(holo) was not observed and obviously loading of the amino acids was also not observed.

4.4.4 Heterologous expression of the A_2-PCP$_2$ didomain

The A_2-PCP$_2$-didomain was cloned into expression vector pET28a to generate the construct pLNA$_2$-PCP$_2$. The expression of N-terminus His-tagged A_2-PCP$_2$ protein was tried in Rosetta (DE3) pLysS with different concentrations of IPTG and with different post induction times and temperatures. But the protein was only identified in the insoluble fraction.

Since the formation of insoluble aggregates could be minimized at low temperatures, protein expression was done at temperatures lower than 16°C using Arctic cells. Arctic competent cells express chaperones for assisting protein folding at lower temperatures. Since the plasmid carrying the genes encoding for chaperones contains a gentamycin resistance cassette, the transformed clones were selected on the LB agar plates with kanamycin and gentamycin. The genetically modified Arctic-pLNA$_2$-PCP$_2$ was then cultivated in LB medium till OD_{600} of 0.5 was reached. The culture was then induced with 0.5mM IPTG and incubated at 11°C for additional 24 hrs. SDS-PAGE of the Ni-NTA purified soluble fractions showed a band of 116 kDa consistent to the molecular weight of A_2-PCP$_2$ protein (Figure 4-65). The protein was desalted and concentrated before proceeding with γ-$^{18}O_4$-ATP pyrophosphate exchange assay.
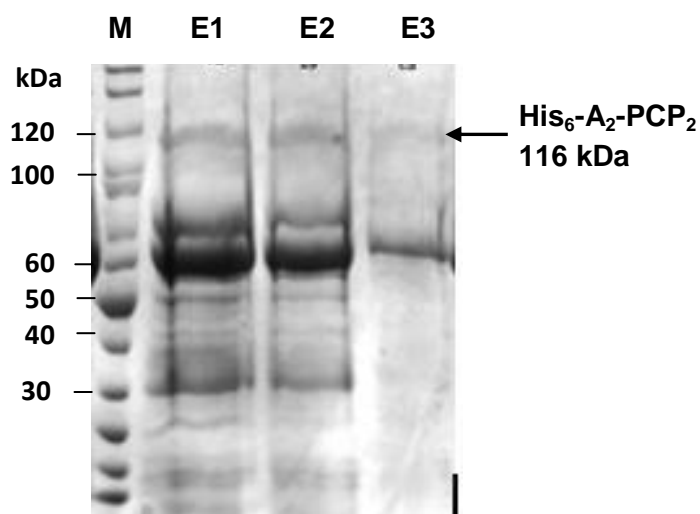
Figure 4-65: Heterologous expression and purification of His$_6$-A_2-PCP$_2$.
SDS-PAGE (10%) of Ni-NTA elution fractions (E1-E3) containing His$_6$-tagged A_2-PCP$_2$ protein. M-Marker.

4.4.4.1 Adenylation assay with A_2-PCP_2 didomain

From the sequence of the A_2 domain the substrate activated by the latter could not yet be predicted. Therefore, A_2-PCP_2 protein was assayed with a broad range of amino acids. Similar to the previous experiments the A_2-PCP_2 didomain protein was highly unspecific and adenylated virtually all amino acids (Figure 4-66).
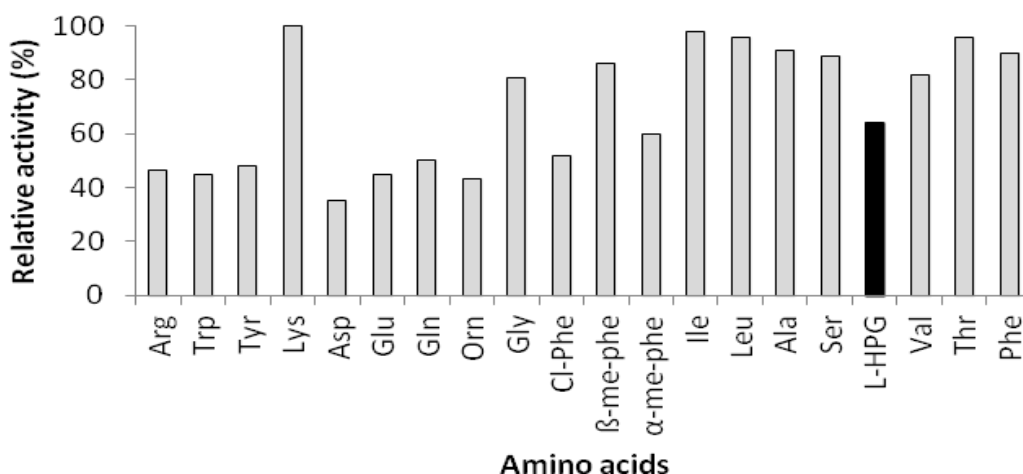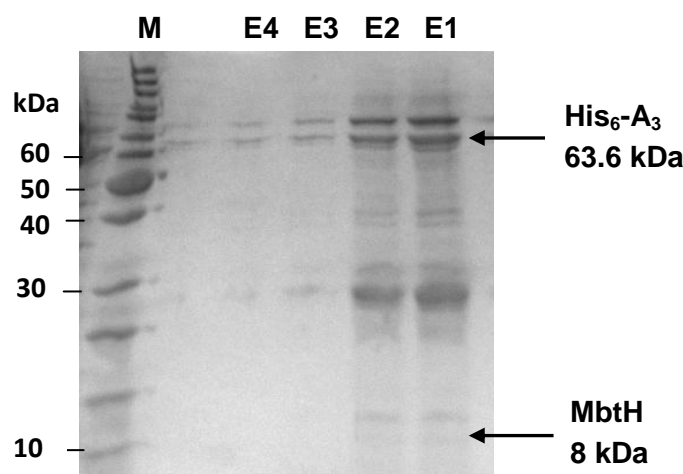

Figure 4-66: γ-$^{18}O_4$-ATP pyrophosphate exchange assay with His_6-tagged A_2-PCP_2 protein. Black bar indicates activity of A_2 domain with the predicted substrate (HPG). Grey bars indicate activity of A_2-PCP_2 domain with other substrates used in the assay.

4.4.5 Coexpression of mbtH-like protein and A_3 domain

The mbtH-like protein is approximately 71 amino acids long and is often encoded in NRPS biosynthetic gene cluster. There are reports claiming the involvement of mbt-H proteins in the adenylation of substrates and regulating the production of non-ribosomal peptides. However, the exact role of mbtH-like proteins is not yet known (Boll et al., 2011; Felnagle et al., 2010). In order to study whether such mbtH-like proteins influence the adenylating proteins A_2, A_3 and A_4 of PKS/NRPS gene locus-1, the *H. aurantiacus* DSM 785 genome was scanned for the presence of *mbtH*-like genes. In total two *mbtH*-like genes were identified at the locus *haur_0026* and *haur_2109*, being located at around 2 Mb and 45 Kb upstream and downstream to PKS/NRPS gene cluster-1, respectively. Because of the relative proximity, the locus *haur_2109* was chosen for coexpression with the adenylating proteins (A_2, A_3 and A_4) of PKS/NRPS gene locus-1.

For the coexpression, initially the *mbtH*-like gene was cloned into the pCDFDuet vector resulting in the construct pLN-*mbtH.* The coexpression vector pCDFDuet was chosen because of its compatibility with the pET28a expression vector.

The pLNA$_3$ and pLN-*mbtH* constructs were transformed into Rosetta (DE3) pLysS expression host. The coexpression of mbtH and A$_3$ domain proteins was attained by the same procedure followed for the expression of discrete A$_3$ domain proteins, except streptomycin was added in addition to kanamycin. SDS-PAGE of the purified protein fractions showed 63.6 kDa band corresponding to the size of the A$_3$ domain protein (Figure 4-67). Additionally another band of 8 kDa consistent with the size of mbtH protein (Figure 4-67) was also identified agreeing with the fact that A domain and mbtH proteins elute together during purification.

Figure 4-67: Heterologous expression and purification of His$_6$-A$_3$-mbtH.
SDS-PAGE (12%) of Ni-NTA elution fractions (E1-E4) containing His$_6$-tagged A$_3$ and mbtH proteins. M, Marker.

4.4.5.1 Adenylation assay with mbtH-A$_3$ domain

The γ-^{18}O$_4$-ATP pyrophosphate exchange assay was done for the expressed mbtH-A$_3$ domain protein in the presence of amino acids used for testing discrete A$_3$ domain. The assay showed higher activation for L-asparagine whereas the other amino acids L-aspartate, L-glutamine, L-cysteine and L-histidine showed very low activation. The other amino acid L-glutamate didn't show any activation. The experiment was repeated again

using L-asparagine along with another set of amino acids, L-threonine, L-tyrosine, L-tryptophan and L-serine. Except L-serine all the other amino acids didn't show any activation (Figure 4-68A). However the activation of L-serine was low when compared to that of L-asparagine. Thus, the earlier experiments with mbtH-A_3 domain protein showed that L-asparagine was the substrate for A_3 domain of PKS/NRPS gene locus-1 (Figure 4-68A). But later on, the experiments with same amino acids showed unspecific activation (Figure 4-68B).

(A)

(B)
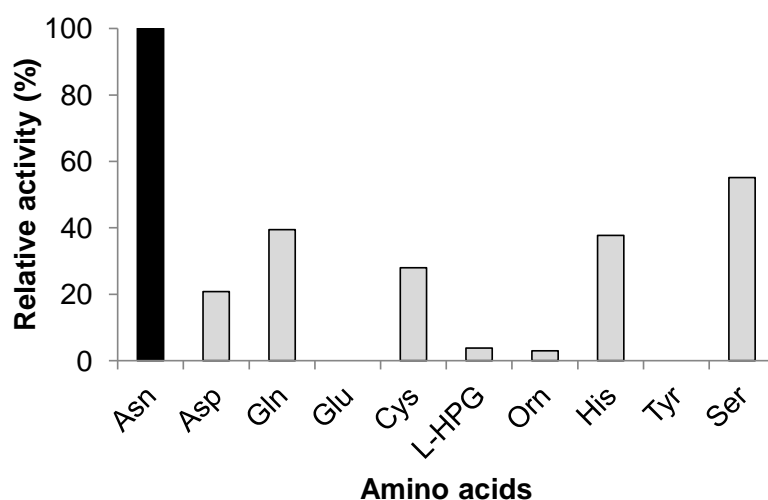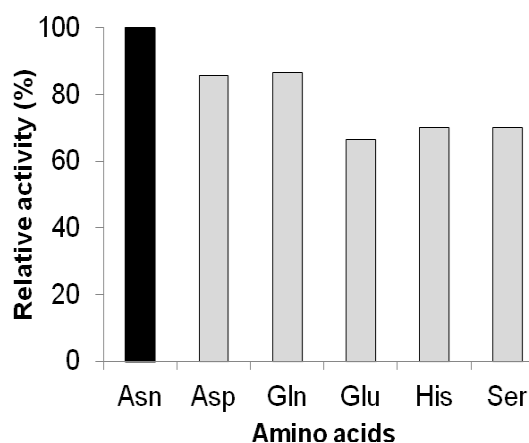
Figure 4-68: (A), (B): γ-^{18}O$_4$-ATP pyrophosphate exchange assay with His$_6$-tagged A$_3$-mbtH protein. Black bar indicates activity of A$_3$ domain with the predicted substrate (Asn). Grey bars indicate activity of A$_3$ domain with other substrates used for the assay.

4.4.6 Coexpression of mbtH-like and A$_4$ domain

As described in the coexpression of mbtH and A$_3$ domain protein, pLN*mbtH* and pLNA$_4$ was cloned and expressed in Rosetta (DE3) pLysS expression host. 49.7 kDa of A$_4$ domain protein and 8 kDa of mbtH-like protein were co-eluted during purification using gravity flow Ni-NTA resin (Figure 4-69).
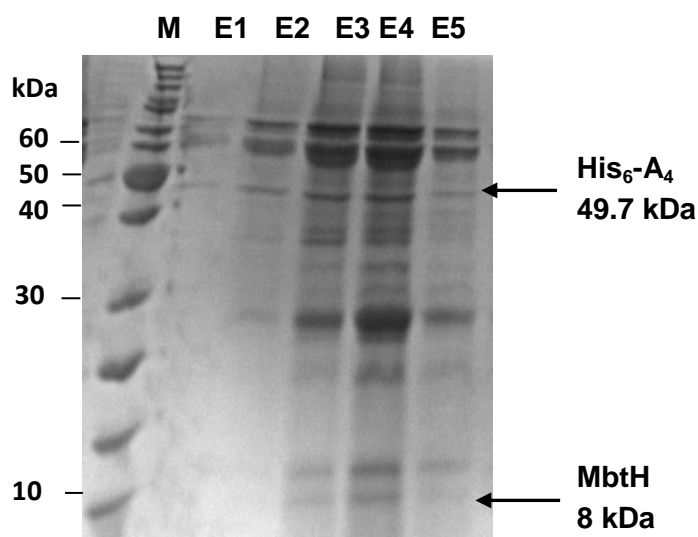
Figure 4-69: Heterologous expression and purification of His$_6$-A$_3$-mbtH.
SDS-PAGE (12%) of Ni-NTA elution fractions (E1-E5) containing His$_6$-tagged A$_4$ and mbtH proteins. M, Marker.

4.4.6.1 Adenylation assay of mbtH-A$_4$ domain

The γ-^{18}O$_4$-ATP pyrophosphate exchange assay was done for the desalted and concentrated mbtH-A$_4$ domain protein with amino acids L-serine, L-threonine, L-cysteine, L-histidine, L-alanine. MALDI-TOF MS analysis of the incubated assay mixture showed activation for L-threonine, whereas the other amino acids tested were not activated (Figure 4-70A). The experiment was repeated with L-threonine along with

additional amino acids L-phenylalanine, L-HPG, L-asparagine, L-glutamine and L-lysine. The assay again showed activation only for L-threonine as expected among the amino acids tested (Figure 4-70A). But, later the γ-$^{18}O_4$-ATP pyrophosphate exchange assay with the same set of amino acids L-serine, L-threonine, L-cysteine, L-histidine and L-valine showed unspecific activation (Figure 4-70B).
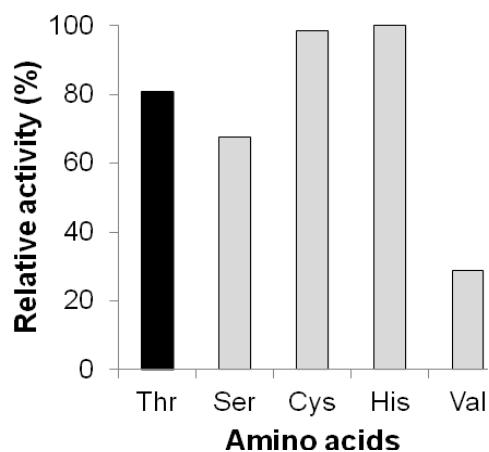
(A)

(B)
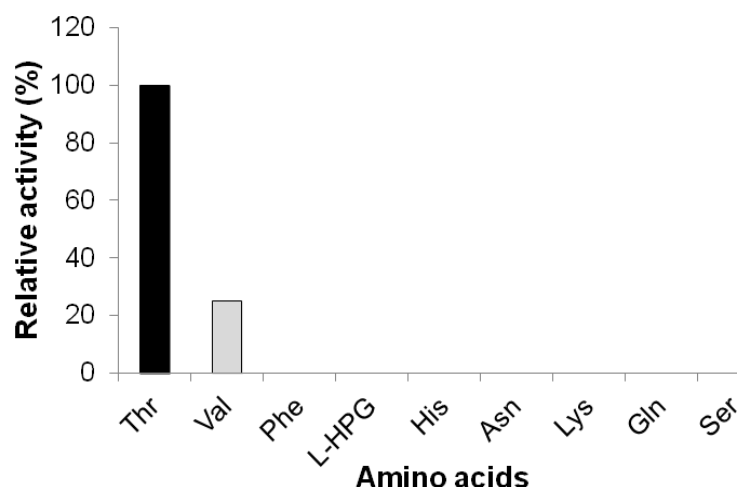
Figure 4-70: (A), (B) γ-$^{18}O_4$-ATP pyrophosphate exchange assay of His_6-tagged A_4-mbtH proteins. Black bar indicates activity of A_4 domain with the predicted substrate (Thr). Grey bars indicate activity of A_4 domain with other substrates used for the assay.

4.4.7 Coexpression and adenylation assay of mbtH-like and A_5 domain

Coexpression of mbtH-like (pLN*mbtH*) and A_5 domain protein (pLNA$_5$) was done in Rosetta (DE3) pLysS expression host using the same procedure as described in section-3.5.5.1. 48.7 kDa of A_5 domain protein and 8 kDa of mbtHe protein were co-eluted during purification using gravity flow Ni-NTA resin (Figure 4-71A). The purified protein was then tested for its adenylation activity using the amino acids L-serine, L-threonine, L-alanine, L-histidine, L-cysteine and L-HPG. MALDI-TOF of the incubated assay mixtures showed the unspecific activation of A_5 domains (Figure 4-71).
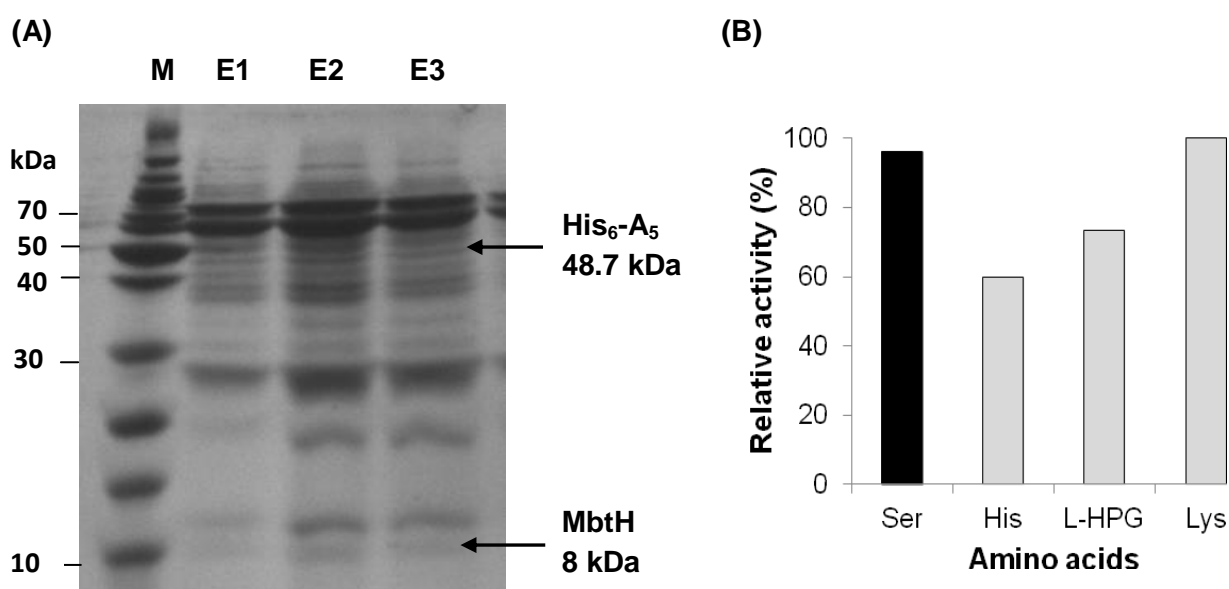

Figure 4-71: γ-$^{18}O_4$-ATP pyrophosphate exchange assay with heterologously expressed and purified His$_6$-A$_5$-mbtH.

(A) SDS-PAGE (12%) of Ni-NTA elution fractions (E1-E3) containing His$_6$-tagged A_5 and mbtH proteins. M, Marker.

(B) γ-$^{18}O_4$-ATP pyrophosphate exchange assay of His$_6$-tagged A_5-mbtH protein. Black bar indicates activity of A_5 domain with the predicted substrate (Ser). Grey bars indicate activity of A_5 domain with other substrates used for the assay.

5. Discussion

Nature is an exquisite resource for structurally diverse molecules. These molecules often exhibit an impressive range of biological properties, which are of pharmacological importance. Since 1981, 41% of clinically approved drugs were either obtained, or inspired by molecules from nature's treasure trove. Of them, 27.4% are being administered as anti-infective (bacterial, fungal, parasitic and viral) drugs and 24.4% are being used for cancer therapies (Newman and Cragg, 2012). Despite the development of natural product based drugs, the existence of untapped organisms in unusual natural habitats and lately especially plentiful genome sequences have made it apparent, that there is a wide range of molecules yet to be determined (Lefevre et al., 2008). The deeper understanding of biosynthetic enzymes has revolutionized the discovery of molecules starting from genes, and offers a unique opportunity for the development of novel pharmaceuticals. For instance, the broad spectrum anti-tumor agent TLN-4601 (ECO-4601), which is currently in phase II clinical trial (McAlpine et al., 2008) was predicted from the genome of the soil *Micromonospora* sp. 046Eco-11, and successively isolated (Bachmann et al., 2004). Almost at the same time TLN-4601 (ECO-4601) was also discovered from the marine *Micromonospora* sp. DPJ12 (Charan et al., 2004).

5.1 Genomes and secondary metabolites of gliding bacteria

The developments in genome-guided discovery and isolation of new compounds is not limited to well studied taxa, but also applicable to the as yet unexplored. In this study we investigated gliding bacterial genomes with a size bigger than 5 Mb, because a genome size bigger than 5 Mb is a typical indicator for the presence of PKS and NRPS genes (Donadio et al., 2007). Subsequently, 33% of the gliding bacterial genomes available through Genbank were recognized by us to harbour PKS and NRPS genes. The striking feature in this is the large number of PKS and NRPS genes identified, which vastly exceeds the number of metabolites isolated from the respective bacteria (Figure 5-1). Reflected by the above, this study sheds light into the latent biosynthetic potential of gliding bacteria, especially from the overlooked genera, as judged from the genomes of *Myxococcus fulvus* HW-1, *Haliangium ochraceum* DSM14365*, Isophaera pallida* ATCC

Discussion

43644, *Paedobacter heparinus* DSM 2366, *Chitinophaga pinensis* DSM 2588 and *Herpetosiphon aurantiacus* DSM 785.
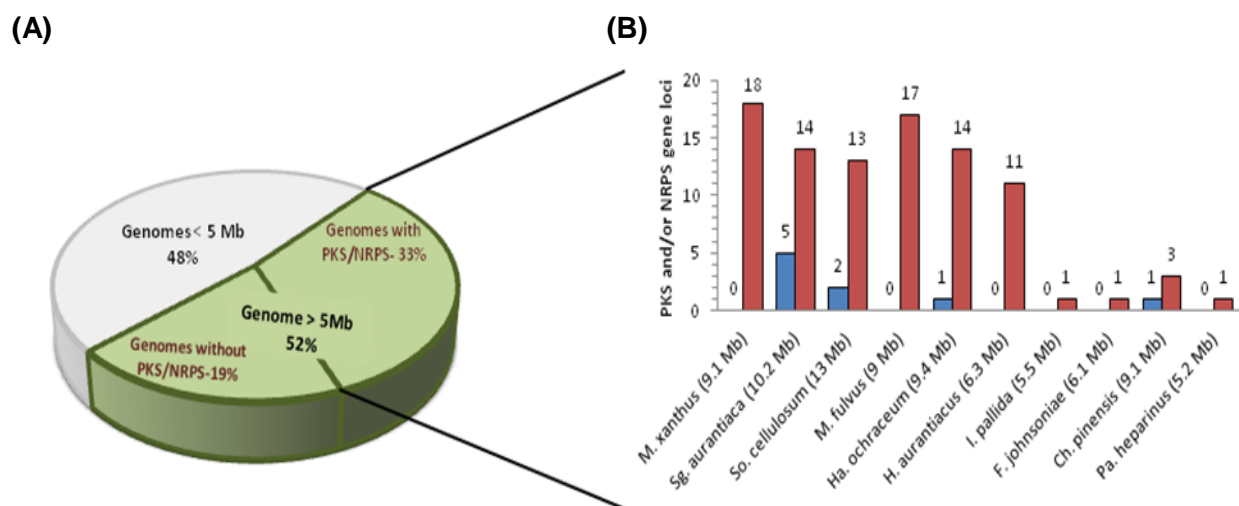
(A) **(B)**

Figure 5-1: Gliding bacterial genomes and their secondary metabolites.
(A) Distribution of gliding bacterial genomes other than cyanobacteria with respect to their size. The percentage of genomes bigger than 5 Mb is subdivided based on the presence of PKS and NRPS genes.
(B) Number of PKS and NRPS genes encoded in gliding bacterial genomes. The red bars represent the genome encoded biosynthetic potential. Blue bars indicate the number of metabolites known from the strains at the time of genome sequencing.

The halotolerant *M. fulvus* HW-1 has 75.17% identical protein coding sequences to those of *M. xanthus* DK1622 (Li et al., 2011) and was expected to have similar biosynthetic abilities. But since the isolation of this strain (Li et al., 2002; Li et al., 2011) no metabolites were reported yet. During our computational screening of the *M. fulvus* HW-1 genome, we identified 17 PKS and/or NRPS gene loci including the putative DKxanthenes and the myxochelin A gene clusters. Since *M. fulvus* HW-1 should be closely related to their terrestrial counterparts (Li et al., 2011), we suggest that at least one of the *M. fulvus* HW-1 cryptic gene clusters encode biosynthetic enzymes for typical *M. fulvus* metabolites (Nett and König, 2007).

With the onset of isolating marine myxobacterial strains, the accessibility of the *Ha. ochraceum* DSM14365 genome affords a possibility to study genes involved in their

secondary metabolism (Ivanova et al., 2010). Only 3% of the genome was found to be involved in the biosynthesis of non-ribosomal peptides and polyketides. Nevertheless, the unusual ecological niche is expected to give rise to the biosynthesis of structurally intriguing compounds. Haliangicin, the antifungal compound isolated from *Ha. ochraceum* DSM14365 (Kundim et al., 2003; Fudou et al., 2001) is an example for this. In our study we rationalized the biosynthetic gene cluster for haliangicin in *Ha. ochraceum* (Section-3.1.1.2.1). However, isotope labeled feeding analysis coupled, either with gene inactivation, or heterologous expression studies are essential for complete validation and identification of the haliangicin biosynthetic pathway. The presence of a methylated myxochelin A biosynthetic pathway in the *Ha. ochraceum* genome shows on the other side that marine myxobacterial strains are closely related to their terrestrial counterparts.

In the smaller genomes of *I. pallida* ATCC 43644 (5.5 Mb) and *Pa. heparinus* DSM 2366 (5.2 MB) one PKS and a hybrid PKS/NRPS gene loci were found, respectively. However, the number of putative PKS and NRPS genes identified is not high, when compared to other genera.

The percentage of PKS and NRPS genes identified from *Ch. pinensis* DSM 2588 (2.9%) is comparable to that from *Ha. ochraceum* DSM 14365. But the size of each of the gene loci is exceptionally big, limiting the number to a single PKS, NRPS and a mixed PKS/NRPS gene locus. The PKS gene locus was already described to involve in the biosynthesis of elansolid B_1 and D (Teta et al., 2010).

At the initial phase of our project only a few genomes of gliding bacteria were publicly accessible. Of them, the *H. aurantiacus* DSM 785 genome was promising because of its size (6 Mb) and lineage. Our detailed bioinformatic analysis of this genome witnessed the presence of 11 PKS and/or NRPS orphan gene loci. Genes to structure prediction of each of these gene loci unfolded the putative structurally most unusual metabolites encoded by the biosynthetic gene cluster. Although no siderophores have yet been isolated from *Herpetosiphon* spp., the identity of one of the mixed PKS/NRPS cluster

with the myxochelin biosynthetic gene cluster, strongly suggests that the resulting siderophore will be myxochelin A.

5.2 Secondary metabolites of *H. aurantiacus* DSM 785

The predictions made from the PKS and NRPS genes of the hitherto unexplored *H. aurantiacus* DSM 785, has prompted us to investigate the associated secondary metabolites of this bacterium. As a consequence of either low, or unexpressed PKS and NRPS genes, however no secondary metabolites were detected from the culture extracts. Therefore, some biosynthetic pathways, i.e. PKS/NRPS gene locus-1, -2 and NRPS gene locus-3 and -4 were examined concerning their expression in the applied cultivation conditions.

RT-PCR analysis of the cryptic biosynthetic gene loci demonstrated the expression of PKS/NRPS gene locus-1 and NRPS gene locus-4 in casitone-yeast (CY) based medium. LC-MS analysis of lyophilised cell extract obtained from CY medium has provided us evidence for the presence of compound-1 and -2. The few micrograms of compound-1 and -2, that are obtained made it possible to perform MS analysis, but this was not sufficient for complete structure determination. According to our structure prediction studies, PKS/NRPS gene locus-1 and NRPS gene locus-4 were determined to encode for a lipohexapeptide and a lipodipeptide, respectively. Since the long chain fatty acid likely to be incorporated by the acyl-CoA synthetase, and the putative substrates activated by some of the A domains could not be predicted with accuracy, the exact molecular mass of the compounds encoded by these gene loci could not be determined. Nevertheless, the putative lipohexapeptide and lipodipeptide were anticipated to have a molecular mass different from that of compound-1 and -2, respectively. So, it has become apparent that although the expression of PKS/NRPS gene locus-1 and NRPS gene locus-4 is ensured, no trace of these lipopeptides could be identified in LC-MS. A reason for this could be that the lipopeptides are produced in amounts too low to be detected.

A possibility to increase the production of the desired metabolites by *H. aurantiacus* DSM 785 would be to alter the cultivation parameters and/or the genetic manipulation of

the bacterium. In this context, it is interesting to note that manipulation of transcriptional regulators has been successful in the up-regulation of the expression of secondary metabolite genes. In *So. cellulosum* So ce 56, overexpression of the positive regulator *chiR* has increased the production of chivosazol (Rachid et al., 2007), whereas inactivation of the negative regulator *ntcA* has enhanced the production of both chivosazol and etnangien (Rachid et al., 2009). With the genome sequence of *H. aurantiacus* DSM 785 in hand, it is possible to identify transcriptional regulators. But the crucial factor is establishing *H. aurantiacus* DSM 785 amenable for genetic manipulation. In our laboratory many different strategies were tried to transform *Herpetosiphon* sp. (Ph.D T. Hoever, 2012), but all rendered unsuccessful.

Heterologous expression of entire cryptic biosynthetic gene loci is the promising alternate route for the production of the corresponding molecules. Since a universal host for heterologous expression is not yet available, the rate-limiting step in this would be the development of a functional genetic system for the expression of *H. aurantiacus* biosynthetic genes. *M. xanthus*, a gliding myxobacteria has been proven successful in expressing a PKS gene cluster encoding for oxytetracycline from the actinomycete, *Streptomyces rimosus* (Stevens et al., 2010). Therefore, *M. xanthus* with a similar lifestyle as that of *Herpetosiphon* would be a challenging host for expressing *H. aurantiacus* biosynthetic genes. Other hosts including *E. coli* and pseudomonads, which has been successful in expressing genes from unrelated organisms (Wenzel and Müller, 2005; Pfeifer et al., 2001) could also be used for the expression of *H. aurantiacus* biosynthetic genes. Additionally, the transfer of biosynthetic genes in a genetically tractable host would give the opportunity to manipulate these genes in such a way that unnatural compounds are generated. The polyketide and anticancer agent epothilone was heterologously expressed in *M. xanthus* and successively the gene cluster was modified to produce novel epothilone analogues (Tang et al., 2005).

Inspite of the breakthroughs in analytical chemistry and molecular biology techniques, the isolation of metabolites encoded by cryptic biosynthetic genes still remains a challenge. Followed by sequencing of the myxobacterial genomes, *Sg. aurantiaca* DW4/3-1 and *So. cellulosum* So ce56, their genomes were found to have 9 cryptic gene

clusters (Schneiker et al., 2007; Ronning and Nierman, 2007). In the span of five years since the identification of these orphan gene clusters, the associated metabolites were not yet isolated. Also, in *M. xanthus* DK1622 six years after sequencing the genome with 13 orphan gene clusters (Wenzel and Müller, 2009b), only one was verified to encode for a new class of compounds called myxoprincomide. The amount of myxoprincomide from *M. xanthus* DK1622 was hardly enough for structure determination. Ultimately, the myxoprincomide gene cluster which was luckily identified in another better producing *M. xanthus* strain was homologously overexpressed under the control of a constitutive promoter (Cortina et al., 2012).

Continuing research in bioinformatics, analytical techniques, molecular biology and synthetic biology will hopefully soon facilitate genome-based discovery of novel compounds. For a big step forward, special microorganisms functioning as platforms for the expression of a wide range of genes are necessary.

5.3 Biochemical characterization of adenylation (A) domains

A domains are the crucial gatekeeper for the entry of building blocks into the construction pipeline of a nonribosomal peptide. Usually A domains exhibit a high degree of substrate specificity. However, A domains with relaxed substrate specificity were also observed thereby giving rise to analogues (Ueki et al., 2006). In this study, we attempted to characterize the A domains of PKS/NRPS gene locus-1, first by bioinformatic analysis and then by the classical γ-$^{18}O_4$- ATP pyrophosphate exchange assay. These methods have been proven accurate for identifying the specific substrate of an A domain (Phelan et al., 2009; Stachelhaus et al., 1999).

Except the A_2 domain, all other A domains, i.e A_3, A_4, A_5, A_7 and A_8 of PKS/NRPS gene locus-1 were found to have a non-ribosomal code at least 90% identical to that of previously characterized A domains (Figure 5-2). It was thus considered most likely that these A domains activate the predicted substrates (Rottig et al., 2011), i.e A_3, A_4 and A_5 are suggested to activate L-asparagine, L-threonine and L-serine, respectively.
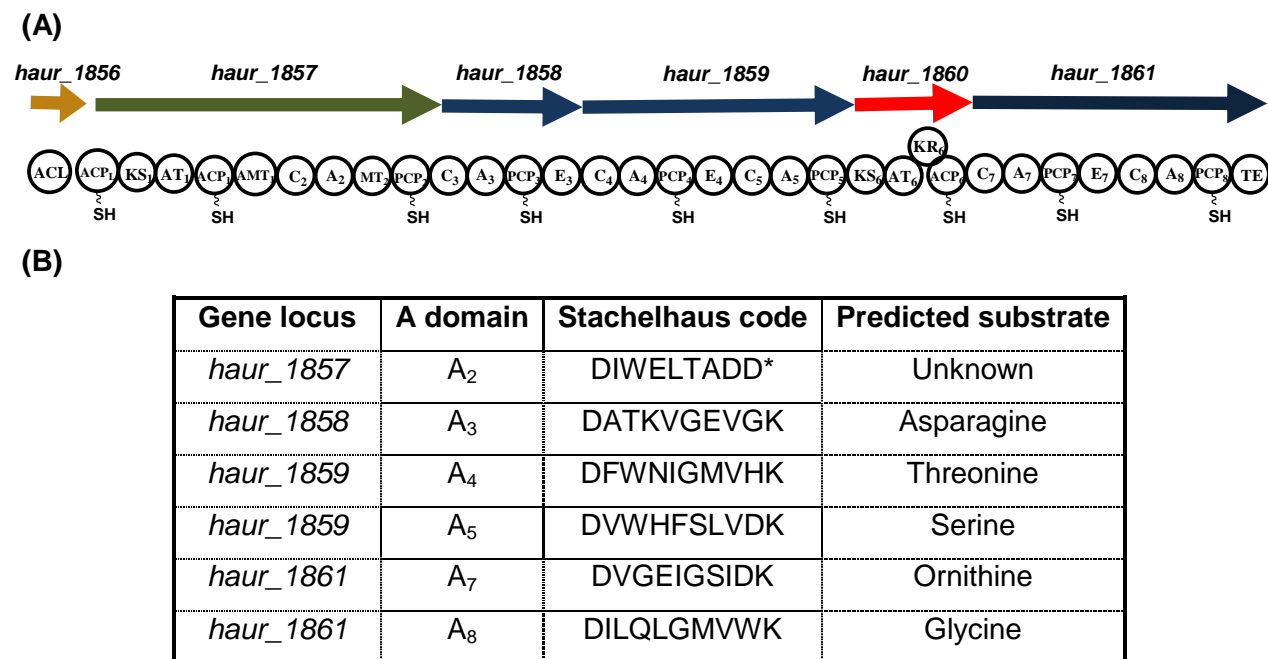
(A)

(B)

Gene locus	A domain	Stachelhaus code	Predicted substrate
haur_1857	A_2	DIWELTADD*	Unknown
haur_1858	A_3	DATKVGEVGK	Asparagine
haur_1859	A_4	DFWNIGMVHK	Threonine
haur_1859	A_5	DVWHFSLVDK	Serine
haur_1861	A_7	DVGEIGSIDK	Ornithine
haur_1861	A_8	DILQLGMVWK	Glycine

Figure 5-2: PKS/NRPS gene locus-1.

(A) NRPS and PKS organization in PKS/NRPS gene locus-1. Putative PKS genes are in red, NRPS genes are blue, PKS/NRPS genes are green and acyl-CoA synthetase is in brown.

(B) A domains of PKS/NRPS gene locus-1 and their substrate specificities.

After heterologously expressing the respective proteins, adenylation assays were carried out in this study with the A_3, A_4 and A_5 domain. However, in contrast to the expected results the proteins showed activation for almost all of the amino acids tested. Moreover, when the same experiment was repeated the results were inconsistent concerning the highly activated amino acid.

It is not uncommon for the A domains with relaxed substrate specificity to behave in vivo contradictory to the computational predictions. For example, one of the A domains in LicC from lichenysin synthetase, activates not only L-isoleucine but also L-valine and L-leucine (Konz et al., 1999). However, this A domain showed 100% identity to the Stachelhaus code for L-isoleucine. In case of xenematide synthase, the Stachelhaus codes for two of the A domains were 70% identical to that of L-phenylalanine. But in vivo, both these A domains activated L-phenylalanine and L-tyrosine thereby producing, xenematide A, B, C and D (Crawford et al., 2011). This clearly demonstrates that

bioinformatics predictions must be viewed critically, and that the biochemistry of especially promiscuous A domains is not yet completely understood.

For *trans* acting A and PCP domains also interdomain interactions have been stated to be important for A domain specificity (Marshall et al., 2002). So, we should not exclude the possibility that interdomain interactions might influence the activity of the A domains, i.e A_3, A_4 and A_5 of PKS/NRPS gene locus-1. To account for possible interdomain interactions the A_2 and A_3 domain of PKS/NRPS gene locus-1 were heterologously expressed together with their PCP domain. But the adenylation assay with A_3-PCP_3 again showed activation of all the tested amino acids at the same rate. The same was observed for the A_2-PCP_2 domain. Thus, it seems apparent that interdomain interactions between A and PCP domains of PKS/NRPS gene locus-1 do not play a role in the activity of these A domains.

For the activity of some A domains a protein called MbtH is important (Felnagle et., 2010; Wolpert et al., 2007), even though the exact function of this protein is not yet reported (Boll et al., 2011). Albeit no *mbtH* gene was found in proximity to PKS/NRPS gene locus-1, the *H. aurantiacus* DSM 785 genome contains two such genes at a distant position. The adenylation assay of the A domains, i.e A_3 and A_4 with the co-expressed mbtH-like proteins showed specific activation of the substrates consistent with the computational prediction. But these results were not reproducible. Moreover, the adenylation assay with A_5-mbtH proteins showed again unspecific activation of amino acids.

The results of the adenylation assays, showing unspecific action of the heterologously expressed A domains raised a lot of questions about the nature of these domains and the reliability of the γ-$^{18}O_4$-ATP pyrophosphate exchange assay. If these A domains were really promiscuous then we presume that *H. aurantiacus* would produce many analogues comprising different combinations of amino acids at a concentration hard to detect via MS. Nevertheless, A domains with such a broad substrate specificity were never observed to date and would be really tempting for further studies. Indeed the γ-$^{18}O_4$-ATP pyrophosphate exchange assay of an A domain activating an amino acid at

higher frequency being not the actual substrate of the biosynthetic process was observed for corallopyronin. The PKS/NRPS corallopyronin cluster encodes a single A domain. This was expected to activate glycine in accordance with the *in silico* analysis and feeding studies. But, in the γ-$^{18}O_4$-ATP pyrophosphate exchange assay with the heterologously expressed corallopyronin A domain, L-threonine was activated approximately two times more than the expected L-glycine. Till date, corallopyronin analogues with L-threonine instead of glycine have not been reported (Erol et al., 2010). It is thus most likely that the γ-$^{18}O_4$-ATP pyrophosphate exchange assay is not really predictive for the here investigated system. Therefore, PKS/NRPS gene locus-1 A domains have to be analyzed for their substrates by assays other than the γ-$^{18}O_4$-ATP pyrophosphate exchange assay.

In NRPS systems, A domains activate their respective substrates by adenylating them at the expense of ATP. This type of adenylation reaction is not only limited to NRPSs but occurs in a wide range of biological processes, including protein synthesis and degradation, DNA synthesis, coenzyme biosynthesis and lipid metabolism. Based on methods used to monitor these adenylation reactions, several non-radioactive assays have been developed to detect the kinetics of pyrophosphate release (McQuade et al., 2009; Otten et al., 2007; Pavela-Vrancicet al., 1999). These non-radioactive assays were successful and can be readily used to distinguish the suitable substrate from other mis-adenylated intermediates. Analysis of PKS/NRPS gene locus-1 A domains of *H. aurantiacus* by these assays would probably give further information about the nature of these peculiar A domains.

To conclude, genome sequences disclose the inherent biosynthetic potential of the respective bacteria. With many ongoing bacterial genome sequencing projects a multitude of biosynthetic pathways will be unveiled in the near future. Along with this, novel technologies will hopefully soon provide the possibility to express these genes heterologously and thus to exploit bacterial metabolomes. Concerning *H. aurantiacus* either the complete heterologous expression of the cryptic gene clusters, or the upregulation of these pathways in the producer strain would have to be attempted to access the encoded natural products.

6. Summary

The need for new drug molecules necessitates different strategies, such as exploring untapped biological sources. Recently, numerous projects focussing bacterial genomes led to new biosynthetic pathways, being useful for drug-oriented natural product research.

In this study, some of the publicly accessible genomes of gliding bacteria were analysed *in silico* using CLUSEAN, PKS/NRPS analysis and NRPS/PKS database for the presence of genes encoding polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). Thus, among 16 completely sequenced genomes from Genbank, *Myxococcus fulvus* HW-1, *Haliangium ochraceum* DSM 14365, *Chitinophaga pinensis* DSM 2588, *Isosphaera pallida* ATCC 43644, *Paedobacter heparinus* DSM 2366 and *Herpetosiphon aurantiacus* DSM 785 were found to contain so called 'cryptic' gene clusters. Such cryptic gene clusters are defined to encode biosynthetic enzymes for hitherto unknown compounds. Deeper understanding regarding the genetic basis and structural enzymology of PKS and NRPS biosynthetic enzymes enable predictions concerning the putative structures encoded by these gene clusters.

M. fulvus HW-1 and *Ha. ochraceum* DSM 14365 both belong to the chemically prolific myxobacteria. In *M. fulvus* HW-1 we identified 15 cryptic PKS and/or NRPS gene loci. Additionally, two gene loci in this bacterial genome were found to be identical to those of the known myxobacterial metabolites DKxanthenes and myxochelin A. *Ha. ochraceum* DSM 14365, the only marine myxobacterium to have its genome sequenced was found to contain 14 PKS and/or NRPS gene loci. Haliangicin, an antifungal compound has already been reported from this bacterium, but the encoding biosynthetic gene cluster is not yet annotated. As a part of this study by taking the haliangicin structure into account, a distinct PKS gene locus was putatively annotated to be associated with its biosynthesis. Again, one of the NRPS gene loci was detected to be homologous to that of the myxochelin A cluster. But, the presence of a methyltransferase encoding gene in proximity to the gene locus suggests that the encoded myxochelin A might be methylated. *Ch. pinensis* DSM 2588 from the phylum Bacteroidetes was already described to produce the polyketides elansolid B_1 and D. Also, a cryptic NRPS gene

locus was reported to be present in the genome sequence. During the current work an additional PKS/NRPS mixed gene locus of 88 Kb was identified from this genome sequence. *I. pallida* ATCC 43644 and *Pa. heparinus* DSM 2366 had been described to be non-producers of secondary metabolites. But, as a part of this study a PKS gene locus was recognized from *I. pallida* ATCC 43644, whereas a PKS/NRPS mixed gene locus was identified from *Pa. heparinus* DSM 2366. The genome of *H. aurantiacus* DSM 785 from the phylum Chloroflexi was identified to putatively own the biosynthetic machineries capable of producing 11 metabolites, including the well known myxochelin A. Thus, from these *in silico* investigations *M. fulvus* HW-1, *Ha. ochraceum* DSM 14365 and *H. aurantiacus* DSM785 were found to be promising candidates for the isolation of novel natural products.

The second part of this study was focussed on *H. aurantiacus* DSM 785. First, using the CLUSEAN, PKS/NRPS analysis, NRPS/PKS database, BLAST, 'NRPSpredictor' and CLUSTALW tools, the putative structures encoded by the cryptic gene clusters of this bacterium were derived. Subsequently, the bacterium was cultivated and screened for the presence of metabolites, whose putative structure had been deciphered from the genome sequence. Since no distinct metabolites could be obtained, the expression of the cryptic gene clusters was studied using RT-PCR. This revealed the expression of PKS/NRPS gene locus-1 and NRPS gene locus-4 in Casitone-yeast based (CY) medium. Cultivation in CY medium and extraction of the lyophilised cell pellet revealed two compounds with molecular masses of 774 Da and 748 Da in LC-MS. These molecular masses of the identified compounds however were unexpected, and their biosynthetic origin could not be correlated with any of the cryptic gene clusters. Further modifications of the culture conditions and of extraction methods did not alter the metabolic profile of the bacterium. Therefore, to succeed in the production of the encoded secondary metabolites either the cryptic gene clusters have to be heterologously expressed or the regulatory genes controlling their expression have to be identified and manipulated.

In the third part of this study we attempted to characterise the putative adenylation (A) domains of PKS/NRPS gene locus-1. A domains are one of the catalytically active

Summary

centres of NRPS and responsible for the selection and adenylation of a specific amino acid at the expense of ATP. Subsequently the adenylated amino acid is loaded on to the peptidyl carrier protein (PCP) domain of NRPS for the assembly of a nonribosomal peptide. PKS/NRPS gene locus-1 includes six A domains for the hexapeptide core of the putative lipopeptide. Bioinformatic analysis (NRPSpredictor) predicted the substrate of five of the six A domains, i.e A_3, A_4, A_5, A_7 and A_8 to be asparagine, threonine, serine, ornithine and glycine, respectively. The A_2 domain was unusual and possibly activates 4-hydroxyphenylglycine. In order to assess the activity of these A domains biochemically, they were heterologously expressed and adenylation of the respective substrates in the presence of other amino acids was monitored using the γ-$^{18}O_4$-ATP pyrophosphate exchange assay. In the first set of experiments discrete A domains were expressed. Then A and PCP domains were expressed together in order to investigate the significance of interdomain interactions in selecting specific amino acids. In a third attempt, the A domains were co-expressed with MbtH proteins. The latter were found to be essential for the activity of some A domains. However, in the γ-$^{18}O_4$-ATP pyrophosphate exchange assay unspecific activation of several amino acids was observed with all these proteins. The outcome of these experiments was unexpected and does not allow to conclude on the structure of the encoded natural product.

Mining bacterial genomes is beneficial in identifying the hidden biosynthetic potential of these species. However, despite the genetic potential the production of the respective secondary metabolites may prove to be difficult. Thus, technologies that enable the heterologous expression of complete biosynthetic gene clusters should be in the focus of future work.

7. References

Ansari, M.Z., Yadav, G., Gokhale, R.S., Mohanty, D., 2004. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res 32, W405–W413.

Bachmann, B.O., 2005. Decoding chemical structures from genomes. Nature Chemical Biology 1, 244–245.

Bachmann, B.O., Ravel, J., 2009. Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. Meth. Enzymol. 458, 181–217.

Balibar, C.J., Vaillancourt, F.H., Walsh, C.T., 2005. Generation of D Amino acid residues in assembly of arthrofactin by dual condensation/epimerization domains. Chemistry & Biology 12, 1189–1200.

Bender, C.L., Palmer, D.A., Peñaloza-Vázquez, A., Rangaswamy, V., Ullrich, M., 1998. Biosynthesis and regulation of coronatine, a non-host-specific phytotoxin produced by Pseudomonas syringae. Subcell. Biochem. 29, 321–341.

Bergendahl, V., Linne, U., Marahiel, M.A., 2002. Mutational analysis of the C-domain in nonribosomal peptide synthesis. European Journal of Biochemistry 269, 620–629.

Bergmann, S., Schümann, J., Scherlach, K., Lange, C., Brakhage, A.A., Hertweck, C., 2007. Genomics-driven discovery of PKS/NRPS hybrid metabolites from Aspergillus nidulans. Nature Chemical Biology 3, 213–217.

Birnboim, H.C., Doly, J., 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res 7, 1513–1523.

Bister, B., Bischoff, D., Nicholson, G.J., Stockert, S., Wink, J., Brunati, C., Donadio, S., Pelzer, S., Wohlleben, W., Süssmuth, R.D., 2003. Bromobalhimycin and Chlorobromobalhimycins—Illuminating the Potential of Halogenases in Glycopeptide Antibiotic Biosyntheses. ChemBioChem 4, 658–662.

Bode, H., Müller, R., 2006. Analysis of myxobacterial secondary metabolism goes molecular. Journal of Industrial Microbiology & Biotechnology 33, 577–588.

Bode, H.B., Müller, R., 2005. The impact of bacterial genomics on natural product research. Angewandte Chemie International Edition 44, 6828–6846.

Bok, J.W., Hoffmeister, D., Maggio-Hall, L.A., Murillo, R., Glasner, J.D., Keller, N.P., 2006. Genomic Mining for Aspergillus Natural Products. Chemistry & Biology 13, 31–37.

Boll, B., Taubitz, T., Heide, L., 2011. Role of MbtH-like proteins in the adenylation of tyrosine during aminocoumarin and vancomycin biosynthesis. J. Biol. Chem. 286, 36281–36290.

Borders, D.B., Francis, N.D., Fantini, A.A., 2006. Extractive purification of lipopeptide antibiotics.

References

Brachmann, A.O., Joyce, S.A., Jenke-Kodama, H., Schwär, G., Clarke, D.J., Bode, H.B., 2007. A Type II polyketide synthase is responsible for anthraquinone biosynthesis in *Photorhabdus luminescens*. ChemBioChem 8, 1721–1728.

Bruner, S.D., Weber, T., Kohli, R.M., Schwarzer, D., Marahiel, M.A., Walsh, C.T., Stubbs, M.T., 2002b. Structural basis for the cyclization of the lipopeptide antibiotic surfactin by the thioesterase domain SrfTE. Structure 10, 301–310.

Caffrey, P., 2003. Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. ChemBioChem 4, 654–657.

Challis, G.L., 2008a. Genome mining for novel natural product discovery. J. Med. Chem. 51, 2618–2628.

Challis, G.L., 2008b. Mining microbial genomes for new natural products and biosynthetic pathways. Microbiology 154, 1555–1569.

Challis, G.L., Ravel, J., 2000. Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. FEMS Microbiology Letters 187, 111–114.

Challis, G.L., Ravel, J., Townsend, C.A., 2000. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chemistry & Biology 7, 211–224.

Chan, Y.A., Boyne, M.T., Podevels, A.M., Klimowicz, A.K., Handelsman, J., Kelleher, N.L., Thomas, M.G., 2006. Hydroxymalonyl-acyl carrier protein (ACP) and aminomalonyl-ACP are two additional type I polyketide synthase extender units. PNAS 103, 14349–14354.

Charan, R.D., Schlingmann, G., Janso, J., Bernan, V., Feng, X., Carter, G.T., 2004. Diazepinomicin, a new antimicrobial alkaloid from a marine Micromonospora sp.†. J. Nat. Prod. 67, 1431–1433.

Chiang, Y.-M., Chang, S.-L., Oakley, B.R., Wang, C.C.C., 2011. Recent advances in awakening silent biosynthetic gene clusters and linking orphan clusters to natural products in microorganisms. Curr Opin Chem Biol 15, 137–143.

Christian, O.E., Compton, J., Christian, K.R., Mooberry, S.L., Valeriote, F.A., Crews, P., 2005. Using Jasplakinolide to turn on pathways that enable the isolation of new chaetoglobosins from *Phomospis asparagi*. J. Nat. Prod. 68, 1592–1597.

Cigler, T., Vahdat, L., 2008. Integrating epothilones into the treatment of patients with metastatic breast cancer: clinical perspectives on incorporating recent data in the practice setting. Clin. Breast Cancer 8 Suppl 4, S166–170.

Conti, E., Stachelhaus, T., Marahiel, M.A., Brick, P., 1997. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. EMBO J 16, 4174–4183.

References

Corre, C., Challis, G.L., 2009. New natural product biosynthetic chemistry discovered by genome mining. Natural Product Reports 26, 977.

Cortina, N.S., Krug, D., Plaza, A., Revermann, O., Müller, R., 2012. Myxoprincomide: A Natural Product from Myxococcus xanthus Discovered by comprehensive analysis of the secondary metabolome. Angewandte Chemie International Edition 51, 811–816.

Crawford, J.M., Portmann, C., Kontnik, R., Walsh, C.T., Clardy, J., 2011. NRPS substrate promiscuity diversifies the xenematides. Org Lett 13, 5144–5147.

Cueto, M., Jensen, P.R., Kauffman, C., Fenical, W., Lobkovsky, E., Clardy, J., 2001. Pestalone, a new antibiotic produced by a marine fungus in response to bacterial challenge. J. Nat. Prod. 64, 1444–1446.

Dagert, M., Ehrlich, S.D., 1979. Prolonged incubation in calcium chloride improves the competence of *Escherichia coli* cells. Gene 6, 23–28.

Dall'Aglio, P., Arthur, C.J., Williams, C., Vasilakis, K., Maple, H.J., Crosby, J., Crump, M.P., Hadfield, A.T., 2011. Analysis of *Streptomyces coelicolor* phosphopantetheinyl transferase, acps, reveals the basis for relaxed substrate specificity. Biochemistry 50, 5704–5717.

Donadio, S., Monciardini, P., Sosio, M., 2007. Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. Natural Product Reports 24, 1073.

Dorrestein, P.C., Blackhall, J., Straight, P.D., Fischbach, M.A., Garneau-Tsodikova, S., Edwards, D.J., McLaughlin, S., Lin, M., Gerwick, W.H., Kolter, R., Walsh, C.T., Kelleher, N.L., 2006. Activity screening of carrier domains within nonribosomal peptide synthetases using complex substrate mixtures and large molecule mass spectrometry. Biochemistry 45, 1537–1546.

Du, L., Lou, L., 2010. PKS and NRPS release mechanisms. Nat Prod Rep 27, 255–278.

Du, L., Shen, B., 2001. Biosynthesis of hybrid peptide-polyketide natural products. Curr Opin Drug Discov Devel 4, 215–228.

Duitman, E.H., Hamoen, L.W., Rembold, M., Venema, G., Seitz, H., Saenger, W., Bernhard, F., Reinhardt, R., Schmidt, M., Ullrich, C., Stein, T., Leenders, F., Vater, J., 1999. The mycosubtilin synthetase of *Bacillus subtilis* ATCC6633: A multifunctional hybrid between a peptide synthetase, an amino transferase, and a fatty acid synthase. Proc Natl Acad Sci U S A 96, 13294–13299.

Erol, Ö., Schäberle, T.F., Schmitz, A., Rachid, S., Gurgui, C., El Omari, M., Lohr, F., Kehraus, S., Piel, J., Müller, R., König, G.M., 2010. Biosynthesis of the myxobacterial antibiotic corallopyronin A. ChemBioChem 11, 1253–1265.

Fazio, G.C., Xu, R., Matsuda, S.P.T., 2004. Genome mining to identify new plant triterpenoids. J. Am. Chem. Soc. 126, 5678–5679.

References

Felnagle, E.A., Barkei, J.J., Park, H., Podevels, A.M., McMahon, M.D., Drott, D.W., Thomas, M.G., 2010. MbtH-like proteins as integral components of bacterial nonribosomal peptide synthetases. Biochemistry 49, 8815–8817.

Finking, R., Marahiel, M.A., 2004. Biosynthesis of nonribosomal peptides. Annual Review of Microbiology 58, 453–488.

Fischbach, M.A., Walsh, C.T., 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem. Rev. 106, 3468–3496.

Fudou, R., Iizuka, T., Sato, S., Ando, T., Shimba, N., Yamanaka, S., 2001. Haliangicin, a novel antifungal metabolite produced by a marine myxobacterium. 2. Isolation and structural elucidation. J. Antibiot. 54, 153–156.

Fudou, R., Jojima, Y., Iizuka, T., Yamanaka, S., 2002. *Haliangium ochraceum* gen. nov., sp. nov. and *Haliangium tepidum* sp. nov.: novel moderately halophilic myxobacteria isolated from coastal saline environments. J. Gen. Appl. Microbiol. 48, 109–116.

Fujimori, D.G., Hrvatin, S., Neumann, C.S., Strieker, M., Marahiel, M.A., Walsh, C.T., 2007. Cloning and characterization of the biosynthetic gene cluster for kutznerides. PNAS 104, 16498–16503.

Gaitatzis, N., Kunze, B., Müller, R., 2001. In vitro reconstitution of the myxochelin biosynthetic machinery of *Stigmatella aurantiaca* Sg a15: Biochemical characterization of a reductive release mechanism from nonribosomal peptide synthetases. PNAS 98, 11136–11141.

Gerth, K., Pradella, S., Perlova, O., Beyer, S., Müller, R., 2003. Myxobacteria: proficient producers of novel natural products with various biological activities—past and future biotechnological aspects with the focus on the genus Sorangium. Journal of Biotechnology 106, 233–253.

Glavina Del Rio, T., Abt, B., Spring, S., Lapidus, A., Nolan, M., Tice, H., Copeland, A., Cheng, J.-F., Chen, F., Bruce, D., Goodwin, L., Pitluck, S., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C.D., Chain, P., Saunders, E., Detter, J.C., Brettin, T., Rohde, M., Göker, M., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C., Klenk, H.-P., Lucas, S., 2010. Complete genome sequence of *Chitinophaga pinensis* type strain (UQM 2034). Stand Genomic Sci 2, 87–95.

Göker, M., Cleland, D., Saunders, E., Lapidus, A., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.-F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Liolios, K., Pagani, I., Ivanova, N., Mavromatis, K., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C.D., Detter, J.C., Beck, B., Woyke, T., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C., Klenk, H.-P., 2011. Complete genome sequence of *Isosphaera pallida* type strain (IS1BT). Stand Genomic Sci 4, 63–71.

Gross et al., 2007. The Genomisotopic Approach: A Systematic Method to Isolate Products of Orphan Biosynthetic Gene Clusters. Chemistry & Biology 14, 53–63.

References

Gulder, T.A.M., Freeman, M.F., Piel, J., 2011. The Catalytic Diversity of Multimodular Polyketide Synthases: Natural Product Biosynthesis Beyond Textbook Assembly Rules. Top Curr Chem.

Gulick, A.M., 2009. Conformational Dynamics in the Acyl-CoA Synthetases, Adenylation Domains of Non-ribosomal Peptide Synthetases, and Firefly Luciferase. ACS Chem. Biol. 4, 811–827.

Hacker, C., Glinski, M., Hornbogen, T., Doller, A., Zocher, R., 2000. Mutational Analysis of the N-Methyltransferase Domain of the Multifunctional Enzyme Enniatin Synthetase. J. Biol. Chem. 275, 30826–30832.

Han, C., Spring, S., Lapidus, A., Del Rio, T.G., Tice, H., Copeland, A., Cheng, J.-F., Lucas, S., Chen, F., Nolan, M., Bruce, D., Goodwin, L., Pitluck, S., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C.C., Saunders, E., Chertkov, O., Brettin, T., Göker, M., Rohde, M., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C., Klenk, H.-P., Detter, J.C., 2009. Complete genome sequence of *Pedobacter heparinus* type strain (HIM 762-3). Stand Genomic Sci 1, 54–62.

Haynes, S.W., Challis, G.L., 2007. Non-linear enzymatic logic in natural product modular mega-synthases and -synthetases. Curr Opin Drug Discov Devel 10, 203–218.

Heathcote, M.L., Staunton, J., Leadlay, P.F., 2001. Role of type II thioesterases: evidence for removal of short acyl chains produced by aberrant decarboxylation of chain extender units. Chemistry & Biology 8, 207–220.

Hertweck, C., 2009. The biosynthetic logic of polyketide diversity. Angewandte Chemie International Edition 48, 4688–4716.

Hubbard, B.K., Thomas, M.G., Walsh, C.T., 2000. Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics. Chemistry & Biology 7, 931–942.

Ikeda, H., Nonomiya, T., Omura, S., 2001. Organization of biosynthetic gene cluster for avermectin in *Streptomyces avermitilis*: analysis of enzymatic domains in four polyketide synthases. J. Ind. Microbiol. Biotechnol. 27, 170–176.

Ivanova, N., Daum, C., Lang, E., Abt, B., Kopitz, M., Saunders, E., Lapidus, A., Lucas, S., Glavina Del Rio, T., Nolan, M., Tice, H., Copeland, A., Cheng, J.-F., Chen, F., Bruce, D., Goodwin, L., Pitluck, S., Mavromatis, K., Pati, A., Mikhailova, N., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C.D., Detter, J.C., Brettin, T., Rohde, M., Göker, M., Bristow, J., Markowitz, V., Eisen, J.A., Hugenholtz, P., Kyrpides, N.C., Klenk, H.-P., 2010. Complete genome sequence of *Haliangium ochraceum* type strain (SMP-2T). Stand Genomic Sci 2, 96–106.

Jarrell, K.F., McBride, M.J., 2008. The surprisingly diverse ways that prokaryotes move. Nature Reviews Microbiology 6, 466–476.

References

Jenke-Kodama, H., Dittmann, E., 2009. Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges. Natural Product Reports 26, 874.

Kalaitzis, J.A., Lauro, F.M., Neilan, B.A., 2009. Mining cyanobacterial genomes for genes encoding complex biosynthetic pathways. Natural Product Reports 26, 1447.

Kastner, S., Müller, S., Natesan, L., König, G., Guthke, R., Nett, M., 2012. 4-Hydroxyphenylglycine biosynthesis in *Herpetosiphon aurantiacus*;: a case of gene duplication and catalytic divergence. Archives of Microbiology 194, 557–566.

Kersten, R.D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C., 2011. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol 7, 794–802.

Kiss, H., Nett, M., Domin, N., Martin, K., Maresca, J.A., Copeland, A., Lapidus, A., Lucas, S., Berry, K., Rio, T.G.D., Dalin, E., Tice, H., Pitluck, S., Richardson, P., Bruce, D., Goodwin, L., Han, C., Detter, J.C., Schmutz, J., Brettin, T., Land, M., Hauser, L., Kyrpides, N.C., Ivanova, N., Göker, M., Woyke, T., Klenk, H.-P., Bryant, D.A., 2011. Complete genome sequence of the filamentous gliding predatory bacterium *Herpetosiphon aurantiacus* type strain (114-95[T]). Stand Genomic Sci 5.

Konz, D., Doekel, S., Marahiel, M.A., 1999. Molecular and biochemical characterization of the protein template controlling biosynthesis of the lipopeptide lichenysin. J. Bacteriol. 181, 133–140.

Konz, D., Marahiel, M.A., 1999. How do peptide synthetases generate structural diversity? Chemistry & Biology 6, R39–R48.

Kopp, F., Marahiel, M.A., 2007. Macrocyclization strategies in polyketide and nonribosomal peptide biosynthesis. Natural Product Reports 24, 735.

Krätzschmar, J., Krause, M., Marahiel, M.A., 1989. Gramicidin S biosynthesis operon containing the structural genes grsA and grsB has an open reading frame encoding a protein homologous to fatty acid thioesterases. J. Bacteriol. 171, 5422–5429.

Krug, D., Zurek, G., Revermann, O., Vos, M., Velicer, G.J., Müller, R., 2008. Discovering the hidden secondary metabolome of *Myxococcus xanthus*: a study of intraspecific diversity. Appl Environ Microbiol 74, 3058–3068.

Kundim, B.A., Itou, Y., Sakagami, Y., Fudou, R., Iizuka, T., Yamanaka, S., Ojika, M., 2003. New haliangicin isomers, potent antifungal metabolites produced by a marine myxobacterium. J. Antibiot. 56, 630–638.

Kunze, B., Reichenbach, H., Müller, R., Höfle, G., 2005. Aurafuron A and B, new bioactive polyketides from *Stigmatella aurantiaca* and *Archangium gephyra* (Myxobacteria). The Journal of Antibiotics 58, 244–251.

Kwan, D.H., Leadlay, P.F., 2010. Mutagenesis of a modular polyketide synthase enoylreductase domain reveals insights into catalysis and stereospecificity. ACS Chem. Biol. 5, 829–838.

References

Kwan, D.H., Schulz, F., 2011. The stereochemistry of complex polyketide biosynthesis by modular polyketide synthases. Molecules 16, 6092–6115.

Kwon, O., Hudspeth, M.E., Meganathan, R., 1996. Anaerobic biosynthesis of enterobactin *Escherichia coli*: regulation of entC gene expression and evidence against its involvement in menaquinone (vitamin K2) biosynthesis. J Bacteriol 178, 3252–3259.

Lai, J.R., Koglin, A., Walsh, C.T., 2006. Carrier protein structure and recognition in polyketide and nonribosomal peptide biosynthesis. Biochemistry 45, 14869–14879.

Lail, K., Sikorski, J., Saunders, E., Lapidus, A., Glavina Del Rio, T., Copeland, A., Tice, H., Cheng, J.-F., Lucas, S., Nolan, M., Bruce, D., Goodwin, L., Pitluck, S., Ivanova, N., Mavromatis, K., Ovchinnikova, G., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C.D., Chain, P., Brettin, T., Detter, J.C., Schütze, A., Rohde, M., Tindall, B.J., Göker, M., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C., Klenk, H.-P., Chen, F., 2010. Complete genome sequence of *Spirosoma linguale* type strain (1T). Stand Genomic Sci 2, 176–185.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Lautru, S., Challis, G.L., 2004. Substrate recognition by nonribosomal peptide synthetase multi-enzymes. Microbiology 150, 1629–1636.

Lautru, S., Deeth, R.J., Bailey, L.M., Challis, G.L., 2005. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. Nature Chemical Biology 1, 265–269.

Lefevre, F., Robe, P., Jarrin, C., Ginolhac, A., Zago, C., Auriol, D., Vogel, T.M., Simonet, P., Nalin, R., 2008. Drugs from hidden bugs: their discovery via untapped resources. Research in Microbiology 159, 153–161.

Li, Y., Müller, R., 2009. Non-modular polyketide synthases in myxobacteria. Phytochemistry 70, 1850–1857.

Li, Y., Weissman, K.J., Müller, R., 2010. Insights into multienzyme docking in hybrid pks–nrps megasynthetases revealed by heterologous expression and genetic engineering. ChemBioChem 11, 1069–1075.

Li, Z.-F., Li, X., Liu, H., Liu, X., Han, K., Wu, Z.-H., Hu, W., Li, F., Li, Y.-Z., 2011. Genome sequence of the halotolerant marine bacterium *Myxococcus fulvus* HW-1. J. Bacteriol. 193, 5015–5016.

Ligon, J., Hill, S., Beck, J., Zirkle, R., Molnár, I., Zawodny, J., Money, S., Schupp, T., 2002. Characterization of the biosynthetic gene cluster for the antifungal polyketide soraphen A from *Sorangium cellulosum* So ce26. Gene 285, 257–267.

Liu, X., Walsh, C.T., 2009. Cyclopiazonic Acid Biosynthesis in Aspergillus sp.: Characterization

References

of a reductase-like R* Domain in cyclopiazonate synthetase that forms and releases cyclo-acetoacetyl-l-tryptophan. Biochemistry 48, 8746–8757.

Liu, X.-Y., Yang, S.-Z., Mu, B.-Z., 2008. Isolation and characterization of a C12-lipopeptide produced by *Bacillus subtilis* HSO 121. Journal of Peptide Science 14, 864–875.

Marahiel, M.A., 2009. Working outside the protein-synthesis rules: insights into non-ribosomal peptide synthesis. Journal of Peptide Science 15, 799–807.

Marahiel, M.A., Stachelhaus, T., Mootz, H.D., 1997. Modular peptide synthetases involved in nonribosomal peptide synthesis. Chem. Rev. 97, 2651–2674.

Marshall, C.G., Burkart, M.D., Meray, R.K., Walsh, C.T., 2002. Carrier protein recognition in siderophore-producing nonribosomal peptide synthetases. Biochemistry 41, 8429–8437.

McBride, M.J., 2001. Bacterial gliding motility: multiple mechanisms for cell movement over surfaces. Annual Review of Microbiology 55, 49–75.

McBride, M.J., Xie, G., Martens, E.C., Lapidus, A., Henrissat, B., Rhodes, R.G., Goltsman, E., Wang, W., Xu, J., Hunnicutt, D.W., Staroscik, A.M., Hoover, T.R., Cheng, Y.-Q., Stein, J.L., 2009. Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. Appl Environ Microbiol 75, 6864–6875.

McClerren, A.L., Cooper, L.E., Quan, C., Thomas, P.M., Kelleher, N.L., van der Donk, W.A., 2006. Discovery and in vitro biosynthesis of haloduracin, a two-component lantibiotic. Proc Natl Acad Sci U S A 103, 17243–17248.

McQuade, T.J., Shallop, A.D., Sheoran, A., DelProposto, J.E., Tsodikov, O.V., Garneau-Tsodikova, S., 2009. A nonradioactive high-throughput assay for screening and characterization of adenylation domains for nonribosomal peptide combinatorial biosynthesis. Analytical Biochemistry 386, 244–250.

Meier, J.L., Burkart, M.D., 2009. The chemical biology of modular biosynthetic enzymes. Chemical Society Reviews 38, 2012.

Meiser, P., Weissman, K.J., Bode, H.B., Krug, D., Dickschat, J.S., Sandmann, A., Müller, R., 2008. DKxanthene biosynthesis—understanding the basis for diversity-oriented synthesis in myxobacterial secondary metabolism. Chemistry & Biology 15, 771–781.

Menche, D., Arikan, F., Perlova, O., Horstmann, N., Ahlbrecht, W., Wenzel, S.C., Jansen, R., Irschik, H., Müller, R., 2008. Stereochemical determination and complex biosynthetic assembly of etnangien, a highly potent rna polymerase inhibitor from the myxobacterium *Sorangium cellulosum*. J. Am. Chem. Soc. 130, 14234–14243.

Mitscher, L., Kunstmann, M., Martin, J., Andres, W., Evans, R., Sax, K., Patterson, E., 1967. Diketopiperazines from fermentations: Metabolites, artifacts, or both. Cellular and Molecular Life Sciences 23, 796–796.

Moore, B.S., Hopke, J.N., 2001. Discovery of a new bacterial polyketide biosynthetic pathway.

References

ChemBioChem 2, 35–38.

Mootz, H.D., Schwarzer, D., Marahiel, M.A., 2002. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. ChemBioChem 3, 490–504.

Mossialos, D., Ochsner, U., Baysse, C., Chablain, P., Pirnay, J.-P., Koedam, N., Budzikiewicz, H., Fernández, D.U., Schäfer, M., Ravel, J., Cornelis, P., 2002. Identification of new, conserved, non-ribosomal peptide synthetases from fluorescent pseudomonads involved in the biosynthesis of the siderophore pyoverdine. Molecular Microbiology 45, 1673–1685.

Naranjo, L., Lamas-Maceiras, M., Ullán, R.V., Campoy, S., Teijeira, F., Casqueiro, J., Martín, J.F., 2005. Characterization of the oat1 gene of *Penicillium chrysogenum* encoding an omega-aminotransferase: induction by L-lysine, L-ornithine and L-arginine and repression by ammonium. Mol. Genet. Genomics 274, 283–294.

Nett, M., Erol, Ö., Kehraus, S., Köck, M., Krick, A., Eguereva, E., Neu, E., König, G.M., 2006. Siphonazole, an unusual metabolite from *Herpetosiphon* sp. Angewandte Chemie International Edition 45, 3863–3867.

Nett, M., Ikeda, H., Moore, B.S., 2009. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Natural Product Reports 26, 1362.

Nett, M., König, G.M., 2007. The chemistry of gliding bacteria. Nat Prod Rep 24, 1245–1261.

Neumann, B., Pospiech, A., Schairer, H.U., 1992. Rapid isolation of genomic DNA from gram-negative bacteria. Trends in Genetics 8, 332–333.

Newman, D.J., Cragg, G.M., 2012. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. J. Nat. Prod. 75, 311–335.

Ohno, A., Ano, T., Shoda, M., 1995. Production of a lipopeptide antibiotic, surfactin, by recombinant *Bacillus subtilis* in solid state fermentation. Biotechnology and Bioengineering 47, 209–214.

Otten, L.G., Schaffer, M.L., Villiers, B.R.M., Stachelhaus, T., Hollfelder, F., 2007. An optimized ATP/PPi-exchange assay in 96-well format for screening of adenylation domains for applications in combinatorial biosynthesis. Biotechnology Journal 2, 232–240.

Pavela-Vrancic, M., Dieckmann, R., Döhren, H.V., Kleinkauf, H., 1999. Editing of non-cognate aminoacyl adenylates by peptide synthetases. Biochem J 342, 715–719.

Pawlik, K., Kotowska, M., Chater, K.F., Kuczek, K., Takano, E., 2007. A cryptic type I polyketide synthase (cpk) gene cluster in *Streptomyces coelicolor* A3(2). Arch. Microbiol. 187, 87–99.

Perlova, O., Gerth, K., Kaiser, O., Hans, A., Müller, R., 2006. Identification and analysis of the chivosazol biosynthetic gene cluster from the myxobacterial model strain *Sorangium cellulosum* So ce56. Journal of Biotechnology 121, 174–191.

References

Pfeifer, B.A., Khosla, C., 2001. Biosynthesis of polyketides in heterologous hosts. Microbiol Mol Biol Rev 65, 106–118.

Phelan, V.V., Du, Y., McLean, J.A., Bachmann, B.O., 2009. Adenylation enzyme characterization using γ -18O4-ATP pyrophosphate exchange. Chemistry & Biology 16, 473–478.

Pradella, S., Hans, A., Spröer, C., Reichenbach, H., Gerth, K., Beyer, S., 2002. Characterisation, genome size and genetic manipulation of the myxobacterium *Sorangium cellulosum*; So ce56. Archives of Microbiology 178, 484–492.

Rachid, S., Gerth, K., Kochems, I., Müller, R., 2007. Deciphering regulatory mechanisms for secondary metabolite production in the myxobacterium *Sorangium cellulosum* So ce56. Molecular Microbiology 63, 1783–1796.

Rachid, S., Gerth, K., Müller, R., 2009. NtcA—A negative regulator of secondary metabolite biosynthesis in *Sorangium cellulosum*. Journal of Biotechnology 140, 135–142.

Rajakumari, S., Daum, G., 2010. Multiple functions as lipase, steryl ester hydrolase, phospholipase, and acyltransferase of tgl4p from the yeast *Saccharomyces cerevisiae*. J Biol Chem 285, 15769–15776.

Rangaswamy, V., Jiralerspong, S., Parry, R., Bender, C.L., 1998. Biosynthesis of the *Pseudomonas* polyketide coronafacic acid requires monofunctional and multifunctional polyketide synthase proteins. Proc Natl Acad Sci U S A 95, 15469–15474.

Rangaswamy, V., Mitchell, R., Ullrich, M., Bender, C., 1998. Analysis of Genes Involved in Biosynthesis of coronafacic acid, the polyketide component of the phytotoxin coronatine. J. Bacteriol. 180, 3330–3338.

Rausch, C., Hoof, I., Weber, T., Wohlleben, W., Huson, D.H., 2007. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evolutionary Biology 7, 78.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., Huson, D.H., 2005. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res 33, 5799–5808.

Richter, C.D., Nietlispach, D., Broadhurst, R.W., Weissman, K.J., 2008. Multienzyme docking in hybrid megasynthetases. Nature Chemical Biology 4, 75–81.

Roberts, A.A., Copp, J.N., Marahiel, M.A., Neilan, B.A., 2009. The *Synechocystis* sp. PCC6803 Sfp-Type phosphopantetheinyl transferase does not possess characteristic broad-range activity. ChemBioChem 10, 1869–1877.

Roongsawang, N., Washio, K., Morikawa, M., 2007. In vivo characterization of tandem C-terminal thioesterase domains in arthrofactin synthetase. ChemBioChem 8, 501–512.

Rottig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., Kohlbacher, O., 2011. NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity.

References

Nucleic Acids Research 39, W362–W367.

Rusnak, F., Sakaitani, M., Drueckhammer, D., Reichert, J., Walsh, C.T., 1991. Biosynthesis of the *Escherichia coli* siderophore enterobactin: sequence of the entF gene, expression and purification of EntF, and analysis of covalent phosphopantetheine. Biochemistry 30, 2916–2927.

Samel, S.A., Marahiel, M.A., Essen, L.-O., 2008. How to tailor non-ribosomal peptide products—new clues about the structures and mechanisms of modifying enzymes. Molecular BioSystems 4, 387.

Sandmann, A., Dickschat, J., Jenke-Kodama, H., Kunze, B., Dittmann, E., Müller, R., 2007. A Type II polyketide synthase from the gram-negative bacterium *Stigmatella aurantiaca* is involved in aurachin alkaloid biosynthesis. Angewandte Chemie International Edition 46, 2712–2716.

Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., et al., 2007. Complete genome sequence of the myxobacterium *Sorangium cellulosum*. Nature Biotechnology 25, 1281–1289.

Schümann, J., Hertweck, C., 2006. Advances in cloning, functional analysis and heterologous expression of fungal polyketide synthase genes. J. Biotechnol. 124, 690–703.

Schwarzer, D., Finking, R., Marahiel, M.A., 2003. Nonribosomal peptides: from genes to products. Nat Prod Rep 20, 275–287.

Schwarzer, D., Marahiel, M.A., 2001. Multimodular biocatalysts for natural product assembly. Naturwissenschaften 88, 93–101.

Schwarzer, D., Mootz, H.D., Linne, U., Marahiel, M.A., 2002. Regeneration of misprimed nonribosomal peptide synthetases by type II thioesterases. Proc Natl Acad Sci U S A 99, 14083–14088.

Sielaff, H., Christiansen, G., Schwecke, T., 2006. Natural products from cyanobacteria: Exploiting a new source for drug discovery. IDrugs 9, 119–127.

Silakowski, B., Kunze, B., Müller, R., 2001. Multiple hybrid polyketide synthase/non-ribosomal peptide synthetase gene clusters in the myxobacterium *Stigmatella aurantiaca*. Gene 275, 233–240.

Silakowski, B., Schairer, H.U., Ehret, H., Kunze, B., Weinig, S., Nordsiek, G., Brandt, P., Blöcker, H., Höfle, G., Beyer, S., Müller, R., 1999. New Lessons for Combinatorial Biosynthesis from Myxobacteria the myxothiazol biosynthetic gene cluster of *Stigmatella aurantiaca* DW4/3-1. J. Biol. Chem. 274, 37391–37399.

Sims, J.W., Schmidt, E.W., 2008. Thioesterase-like role for fungal PKS/NRPS hybrid reductive domains. J. Am. Chem. Soc. 130, 11149–11155.

Smith, S., Tsai, S.-C., 2007. The type I fatty acid and polyketide synthases: a tale of two megasynthases. Nat Prod Rep 24, 1041–1072.

Stachelhaus, T., Mootz, H.D., Marahiel, M.A., 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chemistry & Biology 6, 493–505.

Staunton, J., Weissman, K.J., 2001. Polyketide biosynthesis: a millennium review. Natural Product Reports 18, 380–416.

Steinmetz, H., Gerth, K., Jansen, R., Schläger, N., Dehn, R., Reinecke, S., Kirschning, A., Müller, R., 2011. Elansolid A, a unique macrolide antibiotic from *Chitinophaga sancti* isolated as two stable atropisomers. Angewandte Chemie International Edition 50, 532–536.

Stevens, D.C., Henry, M.R., Murphy, K.A., Boddy, C.N., 2010. Heterologous expression of the oxytetracycline biosynthetic pathway in *Myxococcus xanthus*. Appl Environ Microbiol 76, 2681–2683.

Sudek, S., Haygood, M.G., Youssef, D.T.A., Schmidt, E.W., 2006. Structure of trichamide, a cyclic peptide from the bloom-forming cyanobacterium *Trichodesmium erythraeum*, predicted from the genome sequence. Appl Environ Microbiol 72, 4382–4387.

Tan, L.T., 2007. Bioactive natural products from marine cyanobacteria for drug discovery. Phytochemistry 68, 954–979.

Tang, L., Shah, S., Chung, L., Carney, J., Katz, L., Khosla, C., Julien, B., 2000. Cloning and heterologous expression of the epothilone gene cluster. Science 287, 640–642.

Teta, R., Gurgui, M., Helfrich, E.J.N., Künne, S., Schneider, A., Van Echten-Deckert, G., Mangoni, A., Piel, J., 2010. Genome Mining Reveals trans-at polyketide synthase directed antibiotic biosynthesis in the bacterial phylum Bacteroidetes. ChemBioChem 11, 2506–2512.

Thibodeaux, C.J., Melançon, C.E., Liu, H., 2008. Natural product sugar biosynthesis and enzymatic glycodiversification. Angew Chem Int Ed Engl 47, 9814–9859.

Trick, I., Lingens, F., 1984. Characterization of *Herpetosiphon* sp; —A gliding filamentous bacterium from bulking sludge. Applied Microbiology and Biotechnology 19, 191–198.

Tsai, S.-C. (Sheryl), Ames, B.D., 2009. Structural enzymology of polyketide synthases. Methods Enzymol 459, 17–47.

Ueki, M., Galonić, D.P., Vaillancourt, F.H., Garneau-Tsodikova, S., Yeh, E., Vosburg, D.A., Schroeder, F.C., Osada, H., Walsh, C.T., 2006. Enzymatic generation of the antimetabolite γ,γ-dichloroaminobutyrate by NRPS and mononuclear iron halogenase action in a Streptomycete. Chemistry & Biology 13, 1183–1191.

Valenzano, C.R., Lawson, R.J., Chen, A.Y., Khosla, C., Cane, D.E., 2009. The biochemical basis for stereochemical control in polyketide biosynthesis. J. Am. Chem. Soc. 131, 18501–18511.

References

Vallet-Gely, I., Novikov, A., Augusto, L., Liehl, P., Bolbach, G., Péchy-Tarr, M., Cosson, P., Keel, C., Caroff, M., Lemaitre, B., 2010. Association of hemolytic activity of *Pseudomonas entomophila*, a versatile soil bacterium, with cyclic lipopeptide production. Appl Environ Microbiol 76, 910–921.

Van Lanen, S.G., Shen, B., 2006. Microbial genomics for the improvement of natural product discovery. Current Opinion in Microbiology 9, 252–260.

Van Wagoner, R.M., Drummond, A.K., Wright, J.L.C., 2007. biogenetic diversity of cyanobacterial metabolites: Advances in Applied Microbiology. Academic Press, pp. 89–217.

Walsh, C.T., 2004. Polyketide and nonribosomal peptide antibiotics: modularity and versatility. Science 303, 1805–1810.

Walsh, C.T., 2007. The chemical versatility of natural-product assembly lines. Acc. Chem. Res. 41, 4–10.

Walsh, C.T., Chen, H., Keating, T.A., Hubbard, B.K., Losey, H.C., Luo, L., Marshall, C.G., Miller, D.A., Patel, H.M., 2001. Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. Current Opinion in Chemical Biology 5, 525–534.

Walsh, C.T., Gehring, A.M., Weinreb, P.H., Quadri, L.E., Flugel, R.S., 1997. Post-translational modification of polyketide and nonribosomal peptide synthases. Current Opinion in Chemical Biology 1, 309–315.

Weber, G., Leitner, E., 1994. Disruption of the cyclosporin synthetase gene of *Tolypocladium niveum*. Curr. Genet. 26, 461–467.

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D.H., Wohlleben, W., 2009. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J. Biotechnol. 140, 13–17.

Weissman, K.J., Müller, R., 2009. A brief tour of myxobacterial secondary metabolism. Bioorganic & Medicinal Chemistry 17, 2121–2136.

Weissman, K.J., Müller, R., 2010. Myxobacterial secondary metabolites: bioactivities and modes-of-action. Nat. Prod. Rep. 27, 1276–1295.

Wenzel, S.C., Müller, R., 2005. Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways. Current Opinion in Biotechnology 16, 594–606.

Wenzel, S.C., Müller, R., 2007. Myxobacterial natural product assembly lines: fascinating examples of curious biochemistry. Natural Product Reports 24, 1211.

Wenzel, S.C., Müller, R., 2009a. The impact of genomics on the exploitation of the myxobacterial secondary metabolome. Nat. Prod. Rep. 26, 1385–1407.

References

Wenzel, S.C., Müller, R., 2009b. Myxobacteria—"microbial factories" for the production of bioactive secondary metabolites. Molecular BioSystems 5, 567.

Whitworth, D.E., 2008. Myxobacteria: Multicellularity and Differentiation. ASM Press.

Winter, J.M., Behnken, S., Hertweck, C., 2011. Genomics-inspired discovery of natural products. Current Opinion in Chemical Biology 15, 22–31.

Wolpert, M., Gust, B., Kammerer, B., Heide, L., 2007. Effects of deletions of mbtH-like genes on clorobiocin biosynthesis in *Streptomyces coelicolor*. Microbiology 153, 1413–1423.

Yadav, G., Gokhale, R.S., Mohanty, D., 2003. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. Journal of Molecular Biology 328, 335–363.

Yadav, G., Gokhale, R.S., Mohanty, D., 2009. Towards prediction of metabolic products of polyketide synthases: an in silico analysis. PLoS Comput Biol 5.

Yeh, E., Kohli, R.M., Bruner, S.D., Walsh, C.T., 2004. Type II thioesterase restores activity of a NRPS module stalled with an aminoacyl-s-enzyme that cannot be elongated. ChemBioChem 5, 1290–1293.

Zazopoulos, E., Huang, K., Staffa, A., Liu, W., Bachmann, B.O., Nonaka, K., Ahlert, J., Thorson, J.S., Shen, B., Farnet, C.M., 2003. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nature Biotechnology 21, 187–190.

Zhang, Y.-M., Hurlbert, J., White, S.W., Rock, C.O., 2006. Roles of the active site water, histidine 303, and phenylalanine 396 in the catalytic mechanism of the elongation condensing enzyme of *Streptococcus pneumoniae*. J. Biol. Chem. 281, 17390–17399.

Zheng, J., Keatinge-Clay, A.T., 2011. Structural and functional analysis of C2-type ketoreductases from modular polyketide synthases. Journal of Molecular Biology 410, 105–117.

8. Appendix

Table 8-1: List of genomes analyzed from Genbank. Genomes that were analyzed in this study are highlighted in bold. The annotation of *H. aurantiacus* genome was based on Kiss et al., 2011.

Bacteria	Genome (Mbp)
Phylum: Chloroflexi	
***Chloroflexus aurantiacus* J-10-fl**	**5.3**
***Herpetosiphon auranticaus* DSM 785**	**6**
***Chloroflexus* sp. Y-400-fl**	**5.3**
***Chloroflexus aggregans* DSM 9485**	**4.7**
Phylum: Proteobacteria	
Myxococcus xanthus DK 1622	9.14
Sorangium cellulosum So ce56	13
Stigmatella aurantiaca DW4/3-1	10.2
Anaeromyxobacter dehalogenans 2CP-1-1	5
Anaeromyxobacter sp. K	5
Anaeromyxobacter sp. Fw109-5	5.3
***Haliangium ochraceum* DSM 14365**	**9.4**
***Myxococcus fulvus* HW-1**	
Phylum: Cyanobacteria	
Synechocystis sp. PCC 6803	3.6
Trichodesmium erythraeum IMS101	7.8
Anabaena variabilis ATCC 29413	6.4
Nostoc azollae 0708	5.3
Nostoc punctiforme PCC 73102	8.2
Nostoc sp. PCC 7120	6.4
Phylum: Firmicutes	
Heliobacterium modesticaldum Ice1	*3*
Mycoplasma with	1.3 or less
Phylum: Planctomycetes	
***Isosphaera pallida* ATCC 43644**	**5.5**
Phylum: Bacteroidetes	

Cytophaga hutchinsonii **ATCC 33406**	**4.4**
Capnocytophaga ochracea DSM 7271	2.6
Flavobacterium johnsoniae **UW101**	**6.1**
Flavobacterium psychrophilum JIP02/86	2.9
Cellulophaga algicola **DSM 14237**	**4.9**
Pedobacter heparinus **DSM 2366**	**5.2**
Cytophaga fermentans *DSM 9555*	**6.3**
Spirosoma linguale DSM 74	**8.1**
Chitinophaga pinensis **DSM 2588**	**9.1**
Phylum: Chlorobi	
Chloroherpeton thalassium ATCC 35110	3.3

Figure 8-1: ^{1}H-NMR of fraction containing compound-1 and -2 in methanol-d$_4$
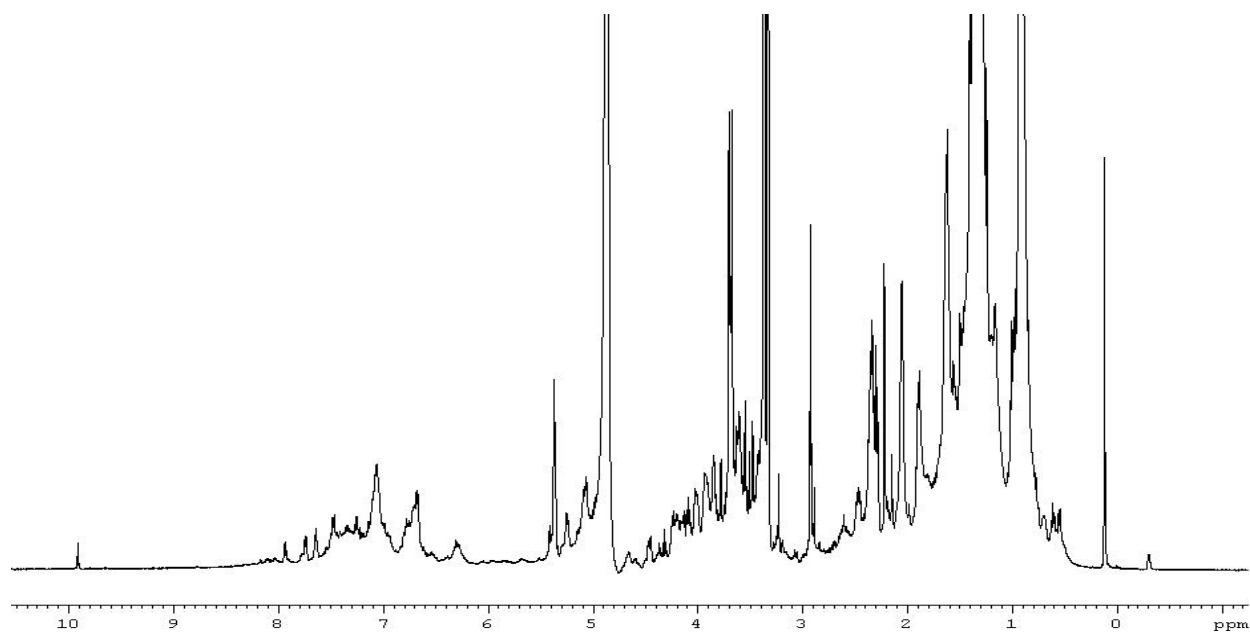
Figure 8-2: ^{13}C-NMR of fraction containing compound-1 and -2 in methanol-d$_4$
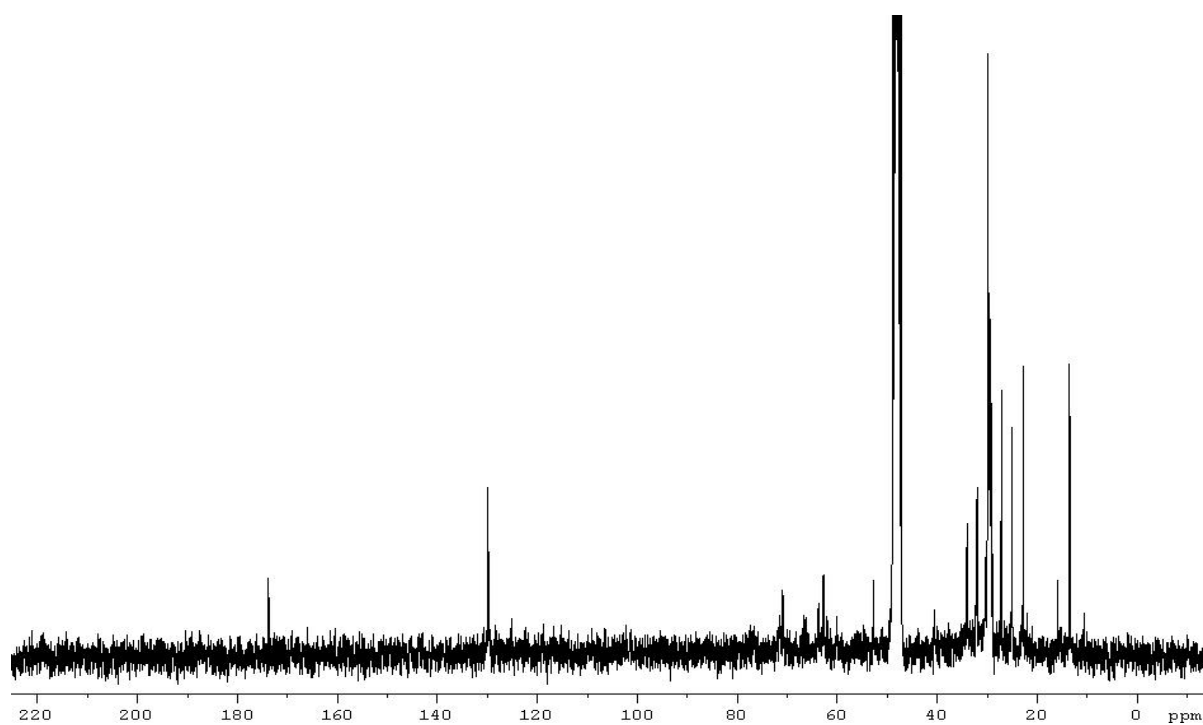
Figure-8.2: LC-MS of fraction containing compound-1 (747 Da) and -2 (773 Da)

(A) Compound-1 (747 Da) in positive (+Q1) and negative (-Q1) mode with DAD spectral data
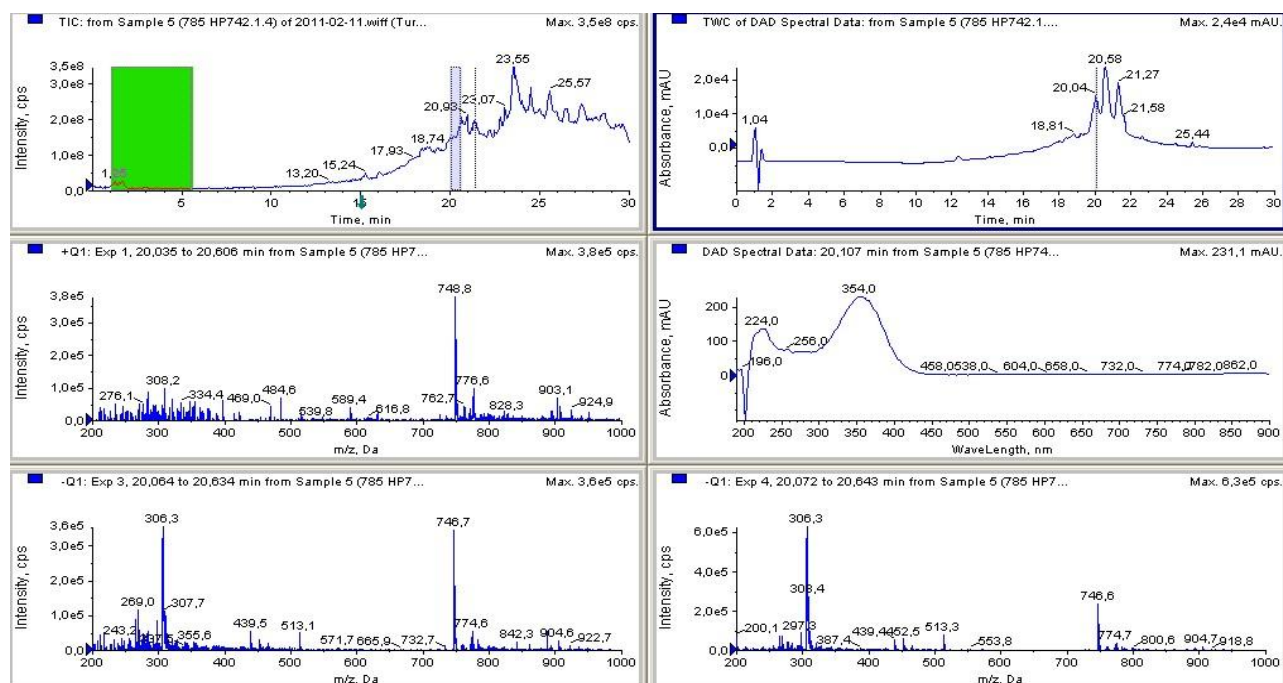
(B) Compound-2 (773 Da) in positive (+Q1) and negative (-Q1) mode with DAD spectral data
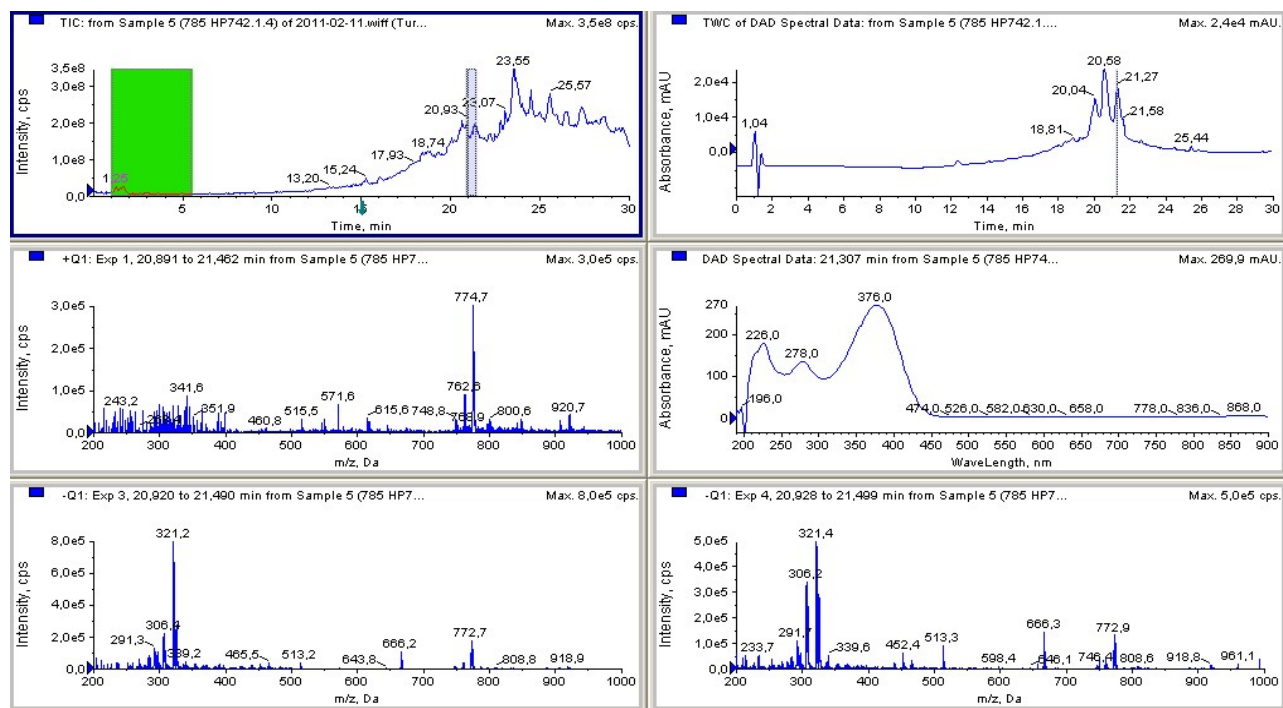
(C) Compound-1 (747 Da) and Compound-2 (773 Da) in positive (+Q1) and negative (-Q1) mode with DAD spectral data
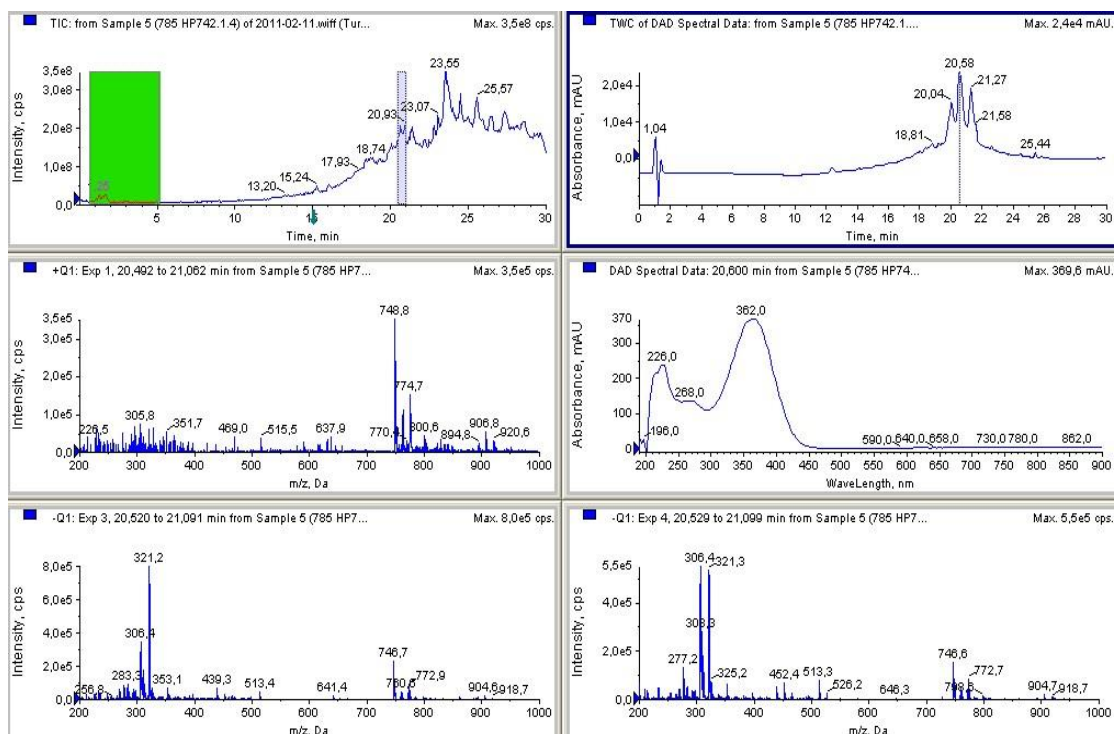
Figure-8.3: HR-MS data of crude extract containg compound with a molecular mass of 772.2198 Da in the positive mode (A) and negative mode (B)
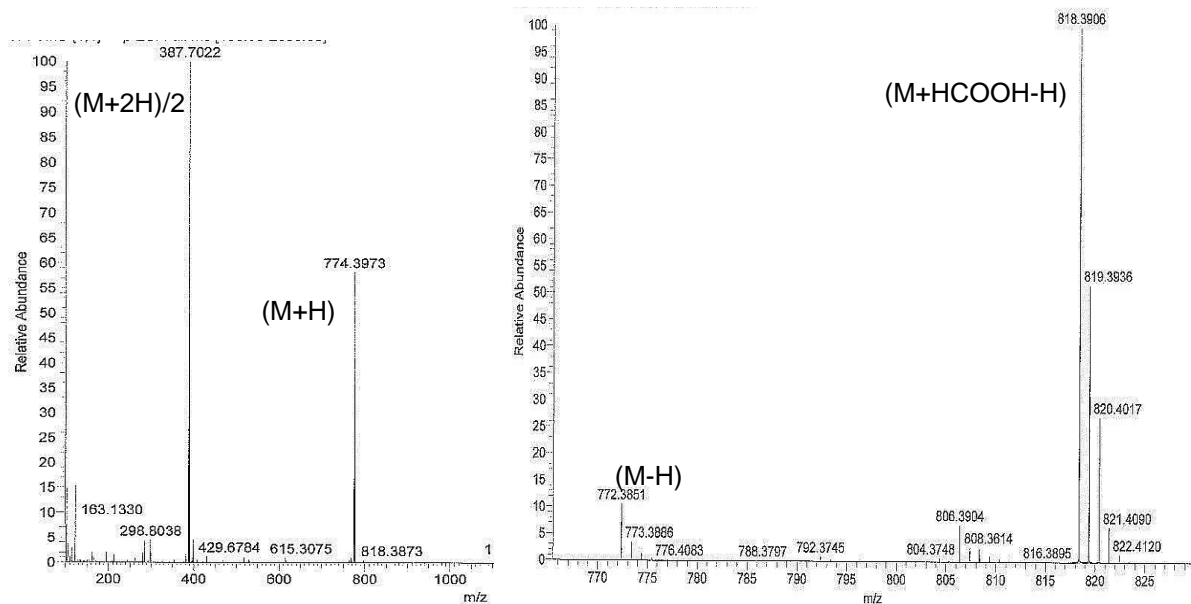
Figure 8-3: Amino acid sequences of overexpressed proteins. His$_6$- six histidine tagged.

Sequences from the pET28-a+ expression vector are underlined and six histidine tag is highlighted in bold.

His$_6$-A$_3$

MGSS**HHHHHH**SSGLVPRGSHMASMTGGQQMGRGSGLVGRLEYNRDLFEPATIARLRDHFLR
LLGHALAQPTQPLAQISILSAAECQQLLVDWNQTQQPFPDQLGLQHLVAQQVQRTPNAPAMR
WNNQIICYTELEQRANQLAHLLLQRGVTQGSIVGVYATRCPEMIISLLAILKAGAAYLPLDPAYPA
ERLHYLVADSAASLIVQASHQALPTLVSTAETLDVVAEAETLASLPTTAPMVDFDPQQLAYVIYT
SGSTGKPKGVLIQHQGVVNYLHWAIHYYPFEQGAGAPLASSLAFDATITALWGPLCTGKTIDLLP
EQDELEVLAQRLSSEDYSVLKITPAHMEALSQLVAPDQIGSSKAFVIGGEALLQQHVAFWQTNA
PNLRLINEYGPTETVVGCVIYQAQAAPSEWAAVPIGRPIANTQLYVLDPAGLPVPIGVPGELYIAG
LGVGRGYHGRPELTAERFVRLEQLAGVQAELARCQQPQPAFERLYRSGDLVRYLPDGNLEYL
GRIDQQVKLHGFRIELGEIEATLASHPTVHAAVAMIREDRPGHKRLVAYVVAEPTANQDTSIVLT
HVAQQLPH

His$_6$-A$_4$

MGSS**HHHHHH**SSGLVPRGSHMSYAELDQRSNQVAHGLIAQGVTVGNLVGLCVERSLELVVGIL
AILKAGAAYVPLDPTYPRERLAFVQADAAIRHIVTQRHLRDVVQAEQCYLLDQPMDAYPTTPPS
VVCSTENPAYVIYTSGSTGNPKGVVVSHANVARLMLATNAWYQFNQHDVWTLFHSYAFDFSV
WELWGALLYGGHLVVVPYWVSRNPEAFHQLLRQQHVTVLNQTPSAFYQLIQADSLAEQRLALR
TVIFGGEALDLAQLAPWFARYGDQQPQLVNMYGITETTVHVTYRPIRLADLQAGLGSVIGCPIPD
LALAVLDAQGRQAGVGVAGELYVGGAGVAQGYLERPELNAQRFIQADASTPDLPSNSRWYRS
GDLVRYWPNGELEYLGRIDLQVKIRGFRIELGEIEAALSQHAAVQSAAV<u>EFELRRQACGRTRAP</u>
<u>PPPPLRSGC</u>

His$_6$-A$_5$

MGSS**HHHHHH**SSGLVPRGSHMSYAELDQRVTQLAASLQAHGVQVDDRVGVLMERSAQLVIAL
LAIVKAGAAYVPFDPAYPSERVLAMLADAAPRVVITDTPKLGQATIPVLLFDQAWQPNHSLSFNP
PIIHPLNAAYMIYTSGSTGKPKGVINSHQAIVNRLLWMQQRYQLTAADVVLQKTPYSFDVSVWE
FFWPLMTGAKLVVARPAGHLDRRYLAETIQAQKVTTIHFVPSMLSLFLEEPQAANCTSLRQVFC
SGEALSAETSARFCQTLNADLHNLYGPTEAAVDVSAWHYQPNAEPSVPIGRPIANTQLYILDAR
MQPVPVGVAGELLIGGLNLARGYAERPDLTAERFIPHPYASQAGARLYRTGDLARWRDDGAIE
YLGRNDFQIKVRGIRVELGEIEHQLSQHPAIAQIVV<u>EFELRRQACGRTRAPPPPPLRSGC</u>

His$_6$-A$_3$-PCP$_3$

MGSS**HHHHHH**SSGLVPRGSHMASMTGGQQMGRGSGLVGRLEYNRDLFEPATIARLRDHFLR
LLGHALAQPTQPLAQISILSAAECQQLLVDWNQTQQPFPDQLGLQHLVAQQVQRTPNAPAMR
WNNQIICYTELEQRANQLAHLLLQRGVTQGSIVGVYATRCPEMIISLLAILKAGAAYLPLDPAYPA
ERLHYLVADSAASLIVQASHQALPTLVSTAETLDVVAEAETLASLPTTAPMVDFDPQQLAYVIYT
SGSTGKPKGVLIQHQGVVNYLHWAIHYYPFEQGAGAPLASSLAFDATITALWGPLCTGKTIDLLP

EQDELEVLAQRLSSEDYSVLKITPAHMEALSQLVAPDQIGSSKAFVIGGEALLQQHVAFWQTNA
PNLRLINEYGPTETVVGCVIYQAQAAPSEWAAVPIGRPIANTQLYVLDPAGLPVPIGVPGELYIAG
LGVGRGYHGRPELTAERFVRLEQLAGVQAELARCQQPQPAFERLYRSGDLVRYLPDGNLEYL
GRIDQQVKLHGFRIELGEIEATLASHPTVHAAVAMIREDRPGHKRLVAYVVAEPTANQDTSIVLT
HVAQQLPHYMLPSVVIWLDSLPLTPNGKVDRQALPAPEINQTALDSAQTTPLDQYEAQLMAIWQ
RVLGLKAVDRHANFFSLGGDSILVMQVVGIARQHGLILTPRLLFQNQTIASLAQAI

His$_6$-A$_2$-PCP$_2$

<u>MGSS**HHHHHH**SSGLVPRGSHMASMTGGQQMGRGSM</u>SYQALDQRANRLANYLQSLAISTNQV
VAILADRSCDFVSAVLGVFKAGAAYLPLDLEHPPRRLAQVLQQSQSRLVLVGEAWQATLAAALS
ILPSDQRPIIVLLEQAFNPELSSEAPTIQSQASDLAYVIYTSGSTGLPKGAMIEQRGMVNHLYAKII
DLQLTAADRVAQNARQSFDISVWQMLVALLVGAETQIYPDSIARDPEVLLSYAEQQATTILEIVPS
LLGAWLTIFPNRANDLPSFAQLRWLLLTGEALPPAACRDWFTWYPTIPLMNAYGPTECSDDVTH
YVVREAPAAHVVHMPIGRPVINTRLYILDGLLQPVPIGVIGELYVGGVGVGRGYLNDPERTQAVF
AADPFMAGGRWYRTGDLARYRSDGTIEYLGRIDHQVKVRGFRIELGEIEAALAQHQAVHQSIVT
ATPNAQGGQLRLIAYVVSKAADQPAEQATSARLEQWDSVWADTYDQLSAGDHGTINTIGWNDSY
TRQPFSAEAMHEWSWVTVRNILAQQPSRILELGCGTGMLLLPLAPYCLSYRGHDIAAEVLAYVQ
QRLDQQIHDWPHVSLAQFPAHDFSNIAPHSVDTIVINSVAQYFPSIDYLVQVLAAGLEALVAGGR
IYLGDIRNLSLNPLLHASIQLFQAPNELAVEQIAQLAQQQHLRDQELVIDPSFFYALQQQYPQISHI
ELVLKRGRIHNELTRFRYDVVLHVQRPSLDLQPHWFDWQADGLSLSLVRQVLEQSQPDALGIA
NVPNSRLSEACGLWQALHVAEQPSTAGELKQALQPLALQGIDPEDWYNLHLNGRYRISVSLAQ
SGELGCYDVLFYATAKLADGGLPQQIQRLASPRKAWSAYANNPLQESQSLTQQFRQHLRQALP
DYMQPEAFVLLEQLPLTPNGKVDRRALAALEAPIQTTTYLAPRNPLEQQLASLFEQVLNLNQVG
VDQSFFELGGHSLTGTQLIGLIRSECHADLPLRTLFEAPTVGELALRVAAAQTEPSEIAKPTALKR
QRQRVNLNTAGFGQIAESNDGGAA

MbtH

MQTTDDDTTIYNVVINHEEQYSIWPAERENPLGWSNVGVSGNKQTCLDHIKAVWTDMRPLSLR
QALANQSA

Figure- 8.4: MALDI-TOF/TOF data of His6-A3-PCP3 (apo) protein. Peptide fragments observed in MALDI-TOF/TOF are highlighted in red. Catalytically active serine (highlighted as ⭐) for the loading of phosphopantetheine arm observed in the digested fragment [HANFFSLGGDSILVMQVVGIARQHGLILTPR] of molecular mass 3346.820 Da

Figure- 8.5: MALDI-TOF MS data of γ-$^{18}O_4$-ATP pyrophosphate exchange assay with His$_6$-A$_3$-mbtH showing specific activitytowards asparagine. The substrates tested and their activation by His$_6$-A$_3$-mbtH protein are indicated in the chromatogram.
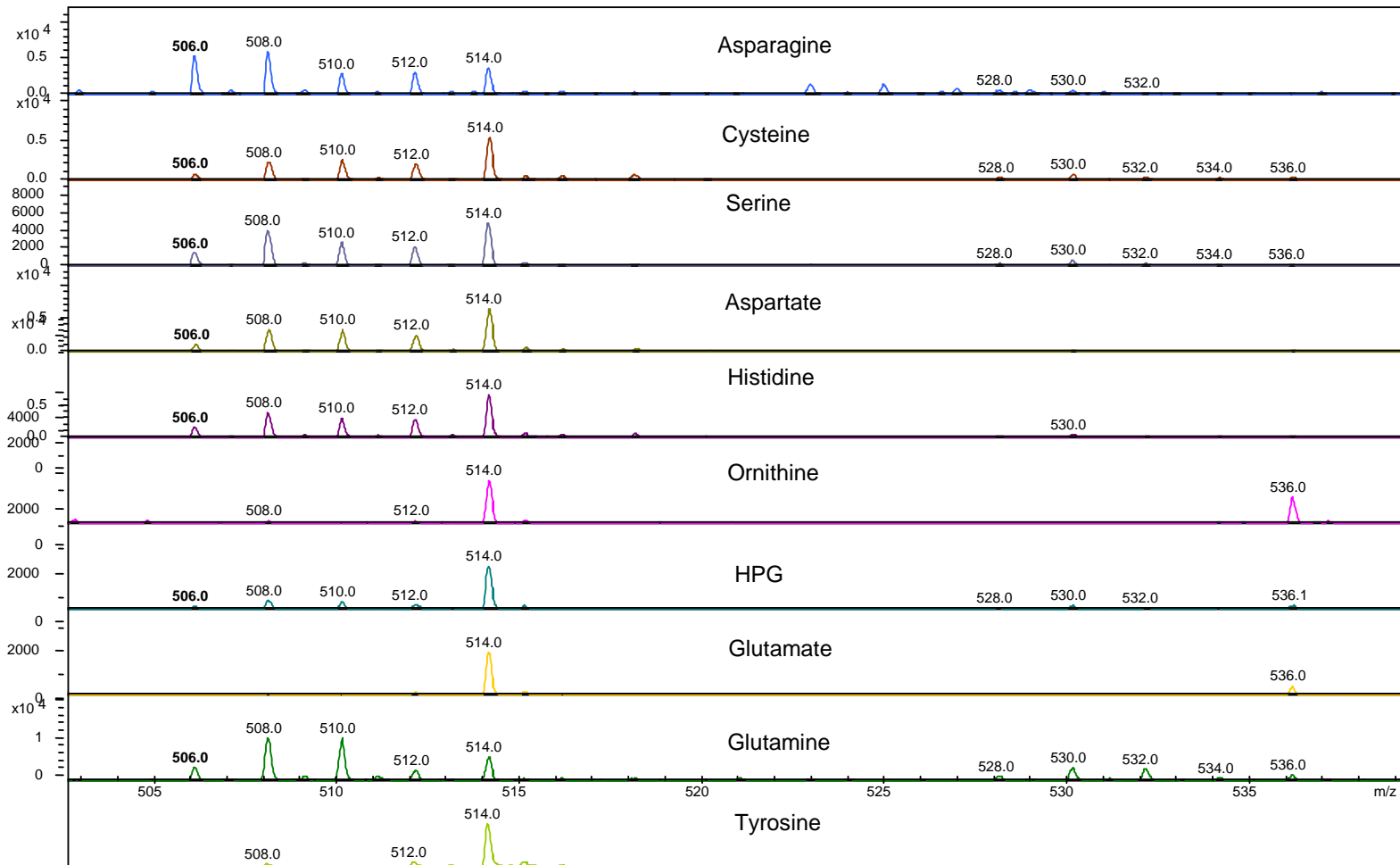
Figure- 8.6: MALDI-TOF MS data of γ-$^{18}O_4$-ATP pyrophosphate exchange assay with His$_6$-A$_3$-mbtH showing unspecific activity. The substrates tested and their activation by His$_6$-A$_3$-mbtH protein are indicated in the chromatogram.